International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 333-343, October-2023

https://doi.org/10.47001/IRJIET/2023.710045

Data Driven Approach to Improve Profitability in Vehicle Insurance Sector

[1]Savindi Welikadaarachchi, [2]Mereesha Botheju, [3]Dinushi Ariyasena, [4]Viruni Fernando, [5]Anjalie Gamage, [6]Poojani Gunathilake

[1,2,3,4,5,6]Department of Information Technology, Sri Lanka Institute of Information Technology, Malabe, Sri Lanka

Authors E-mail: [1]it20134808@my.sliit.lk, [2]it20139230@my.sliit.lk, [3]it20429478@my.sliit.lk, [4]it20187682@my.sliit.lk, [5]anjalie.g@sliit.lk, [6]poojani.g@sliit.lk@sliit.lk

Abstract **- The insurance industry faces significant challenges, including fraudulent claims and customer churn, impacting profitability and sustainability. This research presents a comprehensive data-driven approach to increase profitability in the vehicle insurance sector. This research employs advanced methodologies for Cross sell prediction, Vehicle Insurance Claim Prediction Models, customer survival analysis, churn prediction, and fraud detection using Machine Learning and Deep Learning. By leveraging diverse data sets encompassing policy details, demographics, sentiment analysis, imagery, and historical claims, the study achieves predictive accuracy between 85% to 95%, and fraud detection rates of approximately 80% to 85%. The project introduces the transformative app, VEGO, benefiting both insurance companies and policyholders. Additionally, a rigorous survival analysis addresses critical questions on customer churn dynamics, demonstrating remarkable retention rates. Survival analysis techniques, including Kaplan-Meier Survival Curves, Log-Rank Test, and Cox-proportional hazard models, were employed to analyze customer retention rates over a 72-month period. These models provided valuable insights into risk factors and their cumulative impact on survival time in the insurance context. The research yields insights into risk factors and their collective impact on survival time, while introducing a conservative approach to estimating customer lifetime value. This endeavor enhances analytical foundations in vehicle insurance, ushering in a customer-centric industry landscape.**

Keywords: Data-driven approach, Vehicle insurance sector, Cross sell prediction, Claim prediction models, Customer survival analysis, Churn prediction, Fraud detection, Machine Learning, Deep Learning, Sentiment analysis, Imagery, Predictive accuracy, Kaplan-Meier Survival Curve, Log-Rank Test, Cox-proportional hazard model, Customer lifetime value.

I. INTRODUCTION

The insurance sector is a critical pillar of the global economy, channeling premiums into investments that yield substantial revenue [1]. In Sri Lanka, the insurance industry has witnessed remarkable expansion, contributing significantly to the nation's GDP. Notably, the year 2021 saw an impressive 12.12% surge in Gross Written Premium (GWP), primarily propelled by the popularity and profitability of motor/vehicle insurance [3]. This surge is underpinned by the elevated risks linked to vehicle operation, coupled with the legal mandates for auto insurance in various jurisdictions [4].

In response to evolving dynamics in the insurance landscape, our research delves into a multifaceted approach aimed at leveraging data-driven strategies. This approach encompasses five pivotal components, each poised to revolutionize the vehicle insurance sector

1.1 Cross Sell Prediction

Traditionally reliant on insurance agents for business prospects, we propose an innovative automated machine learning system. This system is designed not only to predict potential life insurance customers but also to facilitate cross-selling opportunities to existing health insurance policyholders [5]. By deploying this strategy, insurers can optimize resources, enhance communication strategies, and bolster revenue streams.

1.2 Fraud Detection

The escalating prevalence of fraudulent claims poses significant financial risks and erodes trust within the insurance industry. In response, our research introduces a comprehensive project, "Detecting Fraudulent Insurance Claims with ML and CNN." This project harnesses the power of machine learning (ML) and convolutional neural networks (CNNs) to revolutionize fraud detection, surpassing traditional methodologies.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 333-343, October-2023

https://doi.org/10.47001/IRJIET/2023.710045

1.3 Claim Prediction

Central to our research is the development of precise predictive models aimed at identifying high-risk policyholders, enabling proactive intervention. This endeavor incorporates deep data preprocessing, advanced feature engineering, and the application of diverse machine learning algorithms, including cutting-edge techniques like Cat Boost and neural networks (MLP, FFNN, RNN).

1.4 Churn Prediction and Customer Lifetime Value Prediction

Recognizing the imperative of retaining valued policyholders, this study delves into churn prediction. Through the application of various techniques, including decision trees, logistic regression, Naive Bayes, random forests, SVM, and survival analysis, the study aims to proactively identify at-risk customers. This allows for focused retention efforts, bolstering long-term customer relationships. An intrinsic element in Customer Relationship Management (CRM), accurate measurement of Customer Lifetime Value (CLV) is indispensable for effective customer management and business growth. This research endeavors to quantify CLV in rupees, providing insurers with invaluable insights into their most valuable clientele.

By addressing these critical components, this research seeks to redefine risk management, customer engagement, and profitability within the dynamic insurance sector.

II. PROBLEM DEFINITION

The insurance industry grapples with challenges like fraudulent claims and customer turnover, affecting profitability and sustainability. This research introduces a data-driven approach to enhance profitability in vehicle insurance. Employing advanced techniques including Machine Learning and Deep Learning, the study achieves predictive accuracy of 85% to 95% and fraud detection rates of 80% to 85%. The project unveils the transformative app, VEGO, benefiting both insurers and policyholders.

Furthermore, rigorous survival analysis delves into customer churn dynamics, revealing impressive retention rates. Techniques such as Kaplan-Meier Survival Curves, Log-Rank Test, and Cox-proportional hazard models were applied, offering insights into risk factors and their cumulative impact on survival time over 72 months. This research not only enhances analytical foundations in vehicle insurance but also introduces a conservative approach to estimating customer lifetime value, ultimately fostering a customer-centric industry landscape.

III. LITERATURE SURVEY

In insurance, companies offer financial support for unforeseen events. Customers choose insurers based on cost and service quality. Recent research highlights decision trees, neural networks, and logistic regression. Babu and Ananthanarayanan's study underscore the prevalence of decision trees and neural networks, particularly in telecommunications. Hybrid models receive less attention in this context.

Table 1: Different Prediction Techniques used in different Studies

Prediction Technique	Reference
Decision Trees	[2]
Logistic Regression	[3][5] [21],
SVM	[4][5][6][7]
Neural Networks	[8]
Genetic Algorithm	[9]
Naïve Bayes	[10][33].
Survival Analysis	[11]
Hazard models	[12]
Decision Trees and Neural Networks	[13][14][9][5]
Logistic Regression and Decision Trees	[7] [14] [15] [16]
Random forest and survival analysis	[15][33].
Decision Trees and Naïve Bayes	[21]
Logistic Regression and Neural Networks	[16][17]
Clustering and Neural Networks	[18]
Decision Trees, Neural Networks and SVM	[19][17]
Decision Trees, Neural Networks, random forest and SVM	[20]

Crawford's investigation in 2015 [3] highlighted the European insurance landscape's notable challenge of grappling with low retention rates. Consequently, insurance enterprises find themselves compelled to sharpen their competitive edge and formulate inventive approaches to cultivate unwavering customer loyalty. This persistent concern regarding customer turnover within the insurance sector has attracted the attention of several scholars [2] [5] [14] [21] [22]. Notably, predictive modeling for customer churn has been a pivotal focus, employing techniques such as neural networks [5] [14] [12], logistic regression [5] [21], and decision trees [7] [14] [15] [16] as the go-to methodologies.

Spiteri's research findings highlight that for predicting customer turnover, Decision Trees (DT), Neural Networks, logistic regression, and Support Vector Machines (SVMs) are frequently utilized tools [28]. However, it's important to note that the specific application and data used may vary between approaches.

ISSN (online): 2581-3048
Volume 7, Issue 10, pp 333-343, October-2023
https://doi.org/10.47001/IRJIET/2023.710045

Among these methods, the Decision Trees (DT) algorithm shines. It excels in handling noisy data, akin to finding patterns in a room full of chatter. Its strength lies in preventing overfitting, a common modeling pitfall, ensuring better generalization to new data. This makes Decision Trees a popular choice in this context.

In Spiteri's study, Neural Networks emerged as the second most favored approach, known for their proficiency in uncovering hidden patterns that simpler models like logistic regression might struggle to identify. This positions neural networks as a go-to solution for various challenges in artificial intelligence.[21] Logistic regression proves invaluable in not only predicting classes but also in highlighting influential factors and their impacts on outcomes [15] Despite the potential power of deep models like Neural Networks, they aren't commonly used for churn-related issues due to the need for substantial data and significant time investment for training [33].

Our study focuses on agent-driven policy churn. While not directly comparing to consumer turnover studies, similar methodologies and techniques are relevant. We define policy churn as adopting a newer policy without clear client benefits [29].

Machine learning and image processing are promising for fraud detection, especially in insurance. XGBoost, used by Najmeddine Dhieb's team, outperforms, achieving 7% higher accuracy in fraudulent claim detection than decision trees [22].

Ismail and Zeadally employed Blockchain (Block-HI) for health insurance fraud detection, analyzing energy-performance trade-offs. More branches slightly improve block execution time, but increased claims significantly impact it [22].

Matloob et al. achieved 85% accuracy in fraud detection, uncovering scenarios missed by manual models. However, handling sensitive data posed challenges due to privacy and integrity concerns [23].

G. Kowshalya and Dr. M. Nandhini utilized RF, J48, and Nave Bayes for classification, potentially benefiting both customers and insurers financially. On the Insurance claim dataset, Random Forest outperformed, while Nave Bayes excelled on the Premium dataset. Further investigations will unveil connections between these datasets, enhancing algorithm performance for real-world applications [15].

Utilizing ML and image processing for insurance fraud detection proves effective. Our project aims to integrate both techniques into a comprehensive model. Previous studies, such as Selvakumar et al. and Alamir, have successfully employed ML for predicting various insurance claims, including motor vehicles and claim statuses [24].Daivi's work encompasses a range of applications including efficiency support for customers, claims fraud detection, expedited claims processing, insurance pricing, individualized suggestions, and customer churn prediction, all using machine learning techniques [25].

Qazi et al. (2017) developed an insurance recommendation system using Bayesian networks, discussed at the Eleventh ACM conference on recommender systems [17].Karp (1998) applied logistic regression to predict customer retention in a financial institution, presented at the Eleventh Northeast SAS Users Group Conference [26].

The insurance dataset is large with potential millions of instances yet exhibits class imbalance. Vaishali Ganganwar (2012) provided an overview of classification algorithms for such imbalanced datasets [27]. Additionally, Peng, Lee, and Ingersoll's 1992 book offers crucial insights into logistic analysis for research applications [28].

A.U. Usman's 2017 research utilized Binary Logistic Regression (LR) for student admissions based on JAMB scores, contributing to model development [29]. Zuriahati Mohd Yunos et al. employed backpropagation neural networks to predict motor insurance claims, considering features like third-party and own damage [30].

In auto insurance, cross-selling involves identifying potential clients for additional services. Methods like acquisition pattern analysis and collaborative filtering (CF), as discussed by Kamakura in 2008, optimize strategies based on prior purchase data [31].

Prinzie & Van den Poel (2011) applied various models, including mixture transition distribution, Markov chain, and Bayesian network, to predict customer behaviors [32].Qazi et al. (2017) developed an insurance recommendation system using Bayesian networks, presented at the Eleventh ACM conference on recommender systems [33].

Li et al. (2011) utilized a multivariate probability model for cross-selling recommendations and later proposed a stochastic dynamic programming model for improved decision-making [34].

Karp, A. (1998), Using logistic regression to predict customer retention, Proceedings of the Eleventh Northeast SAS Users Group Conference. [35]

The insurance data set is characterized by its size and class imbalance. With thousands or even millions of instances, the data set is considered large. However, the distribution of

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 333-343, October-2023

https://doi.org/10.47001/IRJIET/2023.710045

the target variable is heavily skewed, with one class having significantly more instances than the other. Ganganwar, V., 2012. An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering. [36]

Chao-Ying Joanne Peng, Kuk Lida Lee, and Gary M. Ingersoll Have published introduction to logistic analysis book on 1992 which gives a vital knowledge on how logistic regression works on researches [37]

A.U.Usman, 2017 published research on Binary logistic regression analysis on ADDMITING students using jamb score that helps us to analyze on logistic regression to build up the model. [38]

Overall, the literature suggests that data-driven approaches can be effective in improving vehicle insurance profitability. Machine learning algorithms, personalized communication strategies, and NLP techniques can all contribute to better understanding customer behavior and preferences and improve revenue for insurance companies.

IV. METHODOLOGY

The methodology for the development of the predictive models, neural network implementation and customer survival analysis combined with NLP based approach to identify the sentiment of a feedback consisted of the following steps:

Data was collected from an insurance company was preprocessed, transformed for machine learning compatibility, and used to develop models for tasks like cross-selling, fraud detection, and customer value estimation Prior to model development, an exploratory data analysis (EDA) was conducted to gain comprehensive insights into the dataset's characteristics and distributions, providing a crucial foundation for subsequent modeling endeavors.

Training employed Anaconda and Google Colab with scaling, cross-validation, and hyperparameter tuning. Model evaluation utilized precision, recall, and F1 score metrics. A Predictive Pipeline integrated NLP for sentiment analysis in customer retention.

For the integration of these models into practical use, the VEGO application was conceived and developed using Visual Studio Code (VS Code) in conjunction with Python programming. This comprehensive development environment allowed for the seamless amalgamation of various components, ensuring the application's functionality, performance, and user-friendliness. The utilization of Python, a versatile and widely adopted programming language, provided the necessary flexibility and robustness for VEGO's implementation. The integrated workflow of Google Colab for ML and DL model implementation, coupled with thorough EDA, and the utilization of VS Code and Python for app development, collectively constitute the methodological backbone of this research endeavor.

Given below is the high-level architecture diagram of VEGO app.

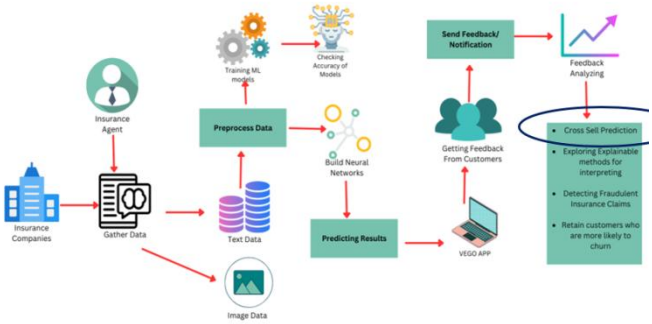

Figure 1: System Architecture

4.1 Cross Sell Prediction

Following meticulous data preprocessing and in-depth analysis, the model-building phase commenced. A diverse array of machine learning algorithms tailored for the intricacies of cross-selling optimization were selected, including Logistic Regression, Gaussian Naive Bayes, Histogram-Based Gradient Boosting, AdaBoost Classifier, Bagging Classifier, Decision Tree, and Random Forest. Deep learning techniques like Recurrent Neural Networks (RNN), Multi-Layer Perceptron (MLP), and Feedforward Neural Network (FFNN) were also incorporated. Additionally, dedicated attention was given to unsupervised learning methods, including K-Nearest Neighbors (KNN) and the XGBoost Classifier. After rigorous experimentation and thorough evaluation of results, the most fitting model was meticulously chosen based on its performance and accuracy. The chosen model's significance and implications will be elaborated upon in the forthcoming Results and Conclusion section, offering a pivotal perspective on cross-selling optimization within this study. The backend creation process encompasses the establishment of a Flask application, the training and serialization of an HGBM model, the development of API endpoints, and the implementation of security measures.

4.2 Claim Prediction

Subsequently, a range of machine learning algorithms including Random Forest, Decision Tree, and Gradient Boosting were implemented using Python and the scikit-learn library, each fine-tuned for optimal performance. Cross-

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 333-343, October-2023

https://doi.org/10.47001/IRJIET/2023.710045

validation techniques were employed to ensure the model's robustness and generalizability. Model evaluation, utilizing metrics such as accuracy, precision, recall, and F1-score, along with techniques like ROC-AUC analysis, provided a comprehensive assessment of its predictive capabilities. Finally, the model was seamlessly integrated into the existing insurance infrastructure for real-time evaluation, demonstrating its practical utility. This rigorous and systematic approach in developing the claim prediction model ensured a reliable tool for anticipating insurance claims..

4.3 Fraud Detection

The first step in building the tabular data analysis model involves gathering comprehensive customer data, past insurance claims, and relevant behavioral information. After a thorough cleaning process to handle missing values and outliers, preprocessing steps like normalization, feature engineering, and categorical variable encoding are applied.

For fraud detection, decision trees and random forests, known for their effectiveness with structured data, are employed. The selected algorithm undergoes training and validation to ensure applicability to unseen data, with potential hyperparameter tuning for optimization.

The image-based fraud detection model focuses on a diverse dataset of vehicle damage images submitted by policyholders. Data augmentation techniques increase diversity, while a tailored CNN architecture extracts features from the images. The model includes Error Level Analysis (ELA) to spot inconsistencies, aiding in the detection of potential fraud.

4.4 Churn Prediction: and customers lifetime value prediction

The proposed system architecture for this section consists of three main components. Customer Churn Prediction, Customer, Survival Analysis and Lifetime Value Prediction and Customer Feedback Analysis and Sentiment Identification

The system calculates customer churn probability and CLV prediction and analyzes customer feedback sentiment using NLP. The backend is developed in Python, with the application built on the Flask framework. Users interact with the system through a user-friendly interface, inputting data for analysis. The UI processes this input, relays it to the backend, and displays churn probability, CLV, and risk level metrics. Additionally, users can select customer feedback to identify its sentiment (positive or negative).

The system relies on a set of tools and technologies, including Visual Studio Code for back-end development, Jupyter Notebook for model training and pipeline construction, Flask as the runtime environment, Python for backend implementation, HTML/bootstrap for the UI, and MySQL for efficient data management.

In summary, this system focuses on predicting customer churn and CLV, as well as analyzing customer feedback sentiments using NLP. Developed in Python with a Flask-based application, it provides an intuitive interface for insurance company analysts to assess critical metrics and quickly identify sentiments in customer feedback.

V. RESULTS AND DISCUSSIONS

5.1 Cross Sell Prediction

The Feed forward Neural Network and Multi-Layer Perceptron models displayed the highest accuracy at 82%, emphasizing their effectiveness in capturing complex patterns. The Gaussian Naïve Bayes also performed well with an accuracy of 80.5%, indicating its aptitude for handling probabilistic relationships. In contrast, models like AdaBoost Classifier, XGBoost Classifier, Decision Tree Classifier, and Random Forest demonstrated lower accuracies (ranging from 50% to 58%) and may benefit from further optimization. The Histogram Based Gradient Boosting Model outshone all others with an impressive 85% accuracy, highlighting its proficiency in capturing intricate patterns. These findings underscore the potential of deep learning and specialized boosting models in this context.

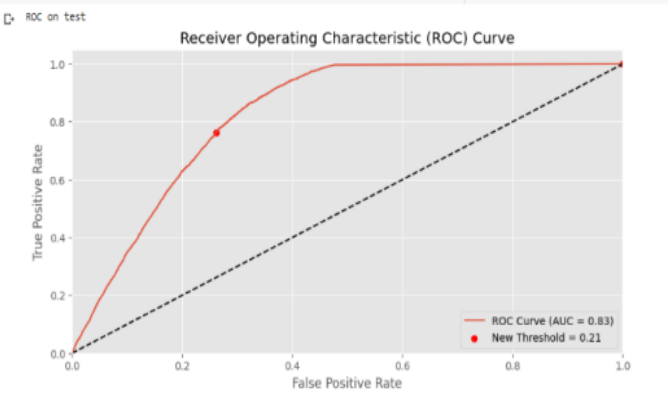

Figure 2: ROC Curve HGBM

SHAP (SHapley Additive exPlanations) is a powerful interpretability technique used in machine learning to understand the contribution of individual features to model predictions, providing valuable insights into the decision-making process of complex models. It quantifies the impact of each feature on prediction and aids in modeling debugging and transparency.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 333-343, October-2023

https://doi.org/10.47001/IRJIET/2023.710045

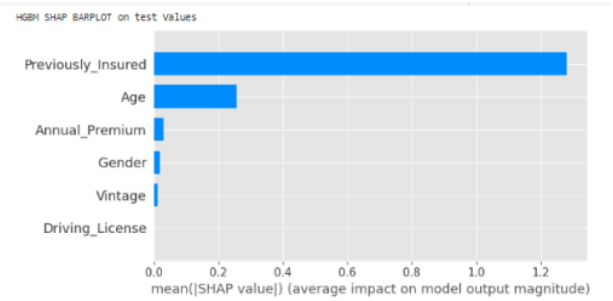

Figure 3: SHAP graph

5.2 Claim Prediction

Through rigorous research, data analysis, and model development, we've made significant strides in enhancing vehicle insurance profitability. A crucial milestone was the meticulous data preprocessing phase, emphasizing the pivotal role of data quality. This involved handling missing values, outliers, and transforming categorical variables to ensure a robust foundation for accurate predictive modeling. Our central achievement lies in the success of our predictive modeling efforts, with the Random Forest Classifier emerging as the top performer, achieving an impressive 87.276% accuracy after thorough hyperparameter tuning. This highlights the practical application of our models in the dynamic insurance industry.

The evaluation of neural network models provided insights beyond accuracy rates. These models' outputs offered detailed probabilistic predictions, enhancing the precision in assessing risk probabilities. This level of granularity in predictions is invaluable for insurance companies, enabling them to tailor promotions and risk mitigation strategies to specific policyholders. Multilayer Perceptrons, Feedforward Neural Networks, and Recurrent Neural Networks provided valuable insights into their capabilities and limitations within this specialized domain.

```
Epoch 96/100
191/191 [==============================] - 0s 2ms/step - loss: 0.0036 - accuracy: 0.9992
Epoch 97/100
191/191 [==============================] - 0s 2ms/step - loss: 0.0050 - accuracy: 0.9988
Epoch 98/100
191/191 [==============================] - 0s 2ms/step - loss: 0.0158 - accuracy: 0.9961
Epoch 99/100
191/191 [==============================] - 0s 2ms/step - loss: 0.0608 - accuracy: 0.9819
Epoch 100/100
191/191 [==============================] - 0s 2ms/step - loss: 0.0346 - accuracy: 0.9886
96/96 [==============================] - 0s 1ms/step - loss: 1.5259 - accuracy: 0.8372
Accuracy: 0.8372246026992798
```

Figure 4: Results for Multilayer Percepton

```
381/381 [==============================] - 1s 2ms/step - loss: 0.2768 - accuracy: 0.8851
Epoch 7/10
381/381 [==============================] - 1s 2ms/step - loss: 0.2590 - accuracy: 0.8915
Epoch 8/10
381/381 [==============================] - 1s 2ms/step - loss: 0.2457 - accuracy: 0.8986
Epoch 9/10
381/381 [==============================] - 1s 2ms/step - loss: 0.2321 - accuracy: 0.9046
Epoch 10/10
381/381 [==============================] - 1s 2ms/step - loss: 0.2131 - accuracy: 0.9144
96/96 [==============================] - 0s 1ms/step - loss: 0.4114 - accuracy: 0.8481
Accuracy: 0.8480762839317322
```

Figure 5: Results for Feedforward Neural Network

```
8/8 [==============================] - 0s 13ms/step - loss: 0.3444 - accuracy: 0.9794
Epoch 7/10
8/8 [==============================] - 0s 14ms/step - loss: 0.2905 - accuracy: 0.9877
Epoch 8/10
8/8 [==============================] - 0s 13ms/step - loss: 0.2402 - accuracy: 0.9959
Epoch 9/10
8/8 [==============================] - 0s 13ms/step - loss: 0.1961 - accuracy: 0.9959
Epoch 10/10
8/8 [==============================] - 0s 14ms/step - loss: 0.1566 - accuracy: 1.0000
2/2 [==============================] - 0s 12ms/step - loss: 0.7892 - accuracy: 0.5500
Accuracy: 0.550000011920929
```

Figure 6: Results for Recurrent Neural Network

5.3 Fraud Detection

Decision Tree Classifier is the first machine learning model used in fraud detection component. The accuracy score is used to assess the overall correctness of the Decision Tree model's predictions. It measures the percentage of correctly classified instances. For the Decision Tree model, the accuracy score is approximately 96.53%, indicating that it correctly predicts fraud or non-fraud claims in nearly 97% of cases.

For the Decision Tree model, the F1-Score is approximately 0.965, the Precision Score is approximately 0.933, and the Recall Score is approximately 1.0. These scores indicate a well-balanced model with high precision and recall.

In the Decision Tree model's confusion matrix, the high count of true positives and true negatives indicates its ability to make correct predictions.

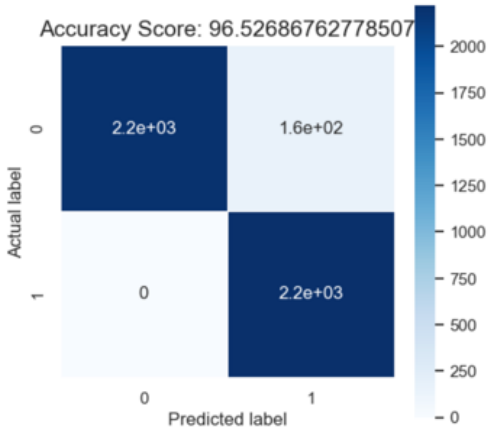

Figure 7: Confusion Matrix for F1 Decision Tree Algorithm

The performance of the model is also measured using the area under the ROC curve (AUC). The AUC for the Decision Tree model is around 0.966, suggesting strong class discrimination.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 333-343, October-2023

https://doi.org/10.47001/IRJIET/2023.710045

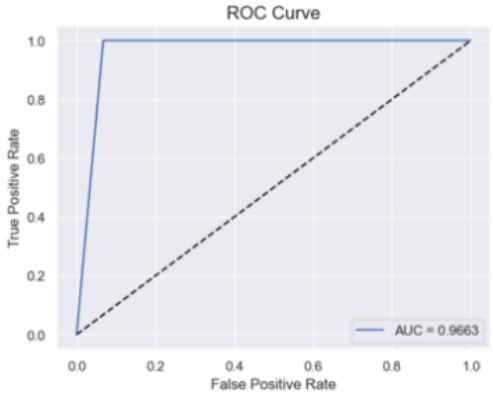

Figure 8: ROC Curve for Decision Tree Algorithm

The Random Forest Classifier, our second machine learning model, achieves an outstanding accuracy of about 99.87%, signifying its exceptional effectiveness in identifying fraudulent claims. Additionally, the model exhibits impressive F1, Precision, and Recall scores, indicating a well-balanced performance with near-perfect precision and recall. The confusion matrix further underscores the model's accuracy, displaying a high count of true positives and true negatives.

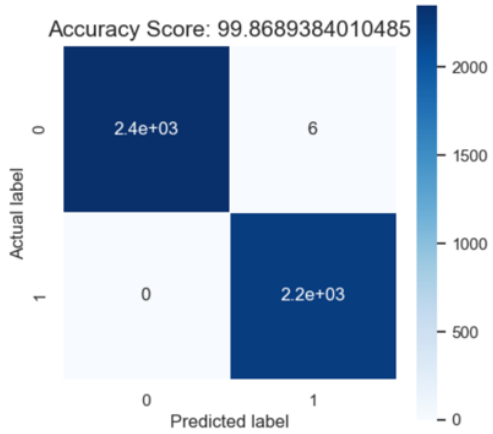

Figure 9: Confusion Matrix for Random Forest Classifier

The ROC curve for the Random Forest model illustrates its superior ability to distinguish between fraudulent and non-fraudulent claims.

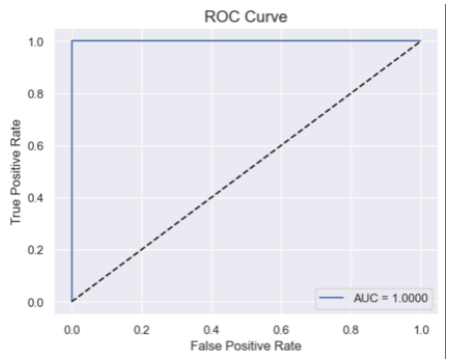

Figure 10: ROC Curve for Random Forest Classifier

In summary, both the Decision Tree and Random Forest models exhibit strong performance in detecting fraudulent insurance claims, as evidenced by high accuracy, precision, recall, F1-Scores, and AUC values.

In Image Based Fraud Detection Model, model.fit method will train the neural network model on the provided training data (X_train, Y_train) for the specified number of epochs. During training, it will also use the validation data (X_val, Y_val) to monitor the model's performance.

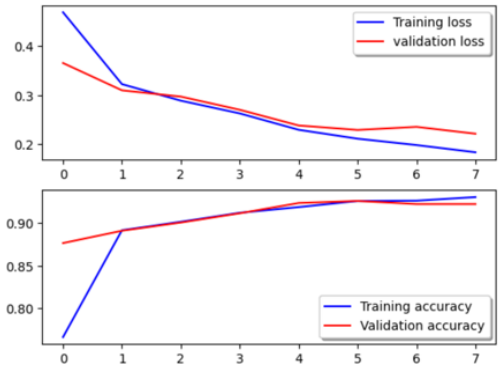

Figure 11: Model Performance

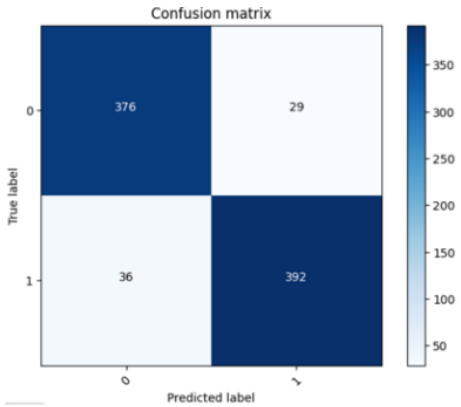

Figure 12: Confusion Matrix for Neural Network

Above charts visualizes the changes in loss and accuracy during training and by visualizing the confusion matrix, which provides insights into how well the model is classifying between authentic and tampered images.

5.4 Churn Prediction: and customers lifetime value prediction

The Random Forest Classifier demonstrated outstanding performance with an accuracy mean of 88.15%, making it the top performer for customer churn prediction in the vehicle insurance sector. Five distinct machine learning and deep learning classifiers were evaluated, considering recall, precision, f1-score, and efficiency due to imbalanced labels. 20% of the dataset was used for testing, while 80% was allocated for training.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 333-343, October-2023

https://doi.org/10.47001/IRJIET/2023.710045

The performance of various classifiers was assessed. Random Forest led with 80.15% accuracy, followed by Decision Tree at 77.03%. Gradient Boost and Adaboost achieved respectable scores of 74.63% and 73.43%, while Voting Classifier scored slightly lower at 71.74%. Among probabilistic models, Naive Bayes achieved 66.41% accuracy, while Kernel SVM and SVM (Linear) scored 66.65% and 66.07% respectively. Logistic Regression had a slightly lower accuracy at 65.74%. K-Nearest Neighbors yielded the lowest accuracy at 58.63%.

A relatively low churn rate in the vehicle insurance sector, with over 60% customer retention after a 70-month period. Notably, there's a consistent decline in survival probability between the 2nd and 58th months. Beyond the 58-month mark, the decline becomes steeper, suggesting a shift in customer retention patterns over time.

The Log-rank test assessed churning probabilities across groups, revealing that gender does not significantly influence survival rates ($p > 0.05$). The null hypothesis suggests no disparity in survival rates between male and female customers, while the alternative hypothesis proposes a significant difference. The p-value of 0.24 indicates a 24% chance of observing a test statistic as extreme or more, assuming the null hypothesis is true. This suggests insufficient evidence to reject the null hypothesis at a standard significance level (e.g., $\alpha = 0.05$), implying the observed survival rate difference may be due to chance.

```
male = (survivaldata['Gender_Male'] == 1)
female = (survivaldata['Gender_Male'] == 0)

plt.figure()
ax = plt.subplot(1,1,1)

kmf.fit(timevar[male],event_observed = eventvar[male],label = "Male")
plot1 = kmf.plot(ax = ax)

kmf.fit(timevar[female],event_observed = eventvar[female],label = "Female")
plot2 = kmf.plot(ax = plot1)

plt.title('Survival of customers: Gender')
plt.xlabel('Tenure')
plt.ylabel('Survival Probability')
plt.yticks(np.linspace(0,1,11))
groups = logrank_test(timevar[male], timevar[female], event_observed_A=eventvar[male], event_observed_B=eventvar[female])
groups.print_summary()
```

Figure 13: Implementing Log Rank test for gender

The Cox-proportional hazard model was used for survival regression, considering multiple risk factors. The hazard rate quantifies the likelihood of the event of interest given survival up to a specific time. The model exhibited a good fit with provided coefficients.

```
cph = CoxPHFitter(penalizer=0.0001)
cph.fit(regression_df, duration_col='tenure', event_col='churn')

cph.print_summary()
```

Figure 14: Implementation of Survival Regression Analysis using Cox Proportional Hazard model

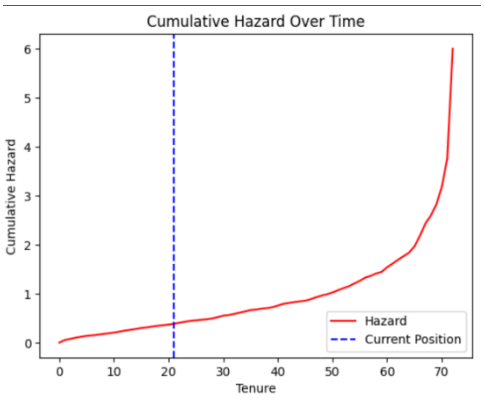

Figure 15: Cumulative Hazard Over time

Figure 16: Explainer which explains the churn risk factors

Figure 17: Gauge chart which shows the churn risk and probability

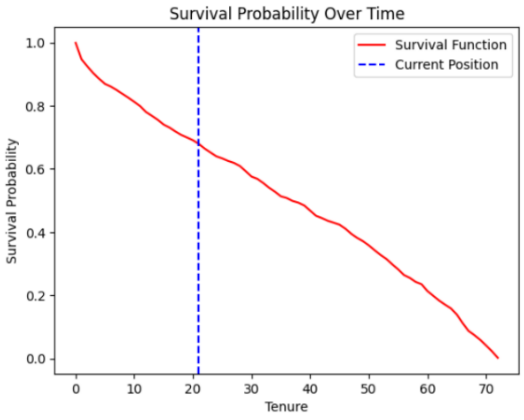

Figure 18: Survival Probability Over Time

The LTV function calculates the Customer Lifetime Value (LTV) using survival analysis predictions. It first predicts the customer's survival function based on provided information. Then, it identifies the maximum tenure where the survival probability exceeds 10%. The LTV is computed by multiplying this maximum tenure by the customer's annual premium. This provides an estimate of the customer's lifetime value, assuming a constant value.

ISSN (online): 2581-3048
Volume 7, Issue 10, pp 333-343, October-2023
https://doi.org/10.47001/IRJIET/2023.710045

```
def LTV(info):
    life = cph.predict_survival_function(info).reset_index()
    life.columns = ['Tenure', 'Probability']
    max_life = life.Tenure[life.Probability > 0.1].max()

    LTV = max_life * info['Annual_Premium '].values[0]
    return LTV

print('LTV of a testid is:', LTV(test_id), 'Rupees.')
```

```
LTV of a testid is: 2506086.0 Rupees.
```

Figure 19: Implementation of Customer Life Time Value Prediction method

After submission of values, four graphics are displayed, including a gauge chart depicting client churn risk, an explainer highlighting key risk factors, and the customer's lifetime value in rupees, along with survival probability and cumulative hazard over time. A multi-tiered risk categorization system categorizes customers into Low, Medium, High, and Extreme risk groups based on churn probabilities. The gauge representation provides an intuitive visual aid for quick risk assessment, employing color-coded segments. This empowers stakeholders to promptly prioritize customers for targeted interventions, enhancing retention efforts.

VI. CONCLUSION

In conclusion, this research has navigated the insurance industry, specifically in the vehicle sector, with a data-driven approach, tackling critical challenges faced by companies. Advanced methodologies applied in cross-sell, claim prediction, customer survival analysis, churn, and fraud detection have yielded impressive results, with accuracy rates ranging from 85% to 95% and fraud detection rates of about 80% to 85%. The introduction of the VEGO app, merging technology and industry insights, promises a transformative interaction between companies and policyholders. Moreover, the study revealed remarkable customer retention rates surpassing 60% even after 72 months, offering deeper insights into the customer base. Additionally, the research sheds light on risk assessment's intricate dynamics, providing companies the tools to enhance their customer-centric approach. Lastly, the innovative fusion of machine learning and convolutional neural networks in combating fraudulent claims not only improves accuracy but also streamlines claims processing, resulting in cost savings and competitive premiums.

ACKNOWLEDGEMENT

This project's successful completion was a collective effort, and we deeply thankful to all those who played a role, whether directly or indirectly. We hope that our work contributes positively to the field of insurance fraud detection and inspires future research endeavors in this domain. Our heartfelt thanks go out to the participants and organizations who provided access to the data and resources necessary for this project. Your cooperation and willingness to share information were essential in enabling us to conduct meaningful research. Finally, we express our gratitude to the broader academic and research community for their valuable contributions to the field of machine learning. Your work served as a constant source of inspiration and knowledge throughout this project.

REFERENCES

[1] H. Z. a. P. Z. Y. Chen, *Study of Customer Lifetime Value Model Based on Survival-Analysis Methods*, pp. 266-270, 2009.

[2] B. M. D. M. a. B. L. P. Datta, "Automated cellular modeling and prediction on a large scale"," vol. 14, pp. 485-502, 2000.

[3] S. H. a. Y. L. J. Ahn, "Customer churn analysis: Churn determinants and mediation effects of partial defection in the korean mobile telecommunications service industry," vol. 30, pp. 552-568, 2006.

[4] V. R. a. V. S. G. G. Sundarkumar, "One-class support vector machine based undersampling," pp. 1-7, 2015.

[5] I. B. a. G. Toderean, "Churn prediction in the telecommunications sector using support vector machine," vol. 22, pp. 1-5, 2013.

[6] L. Y. a. X. Guo-en, "he explanation of support vector machine in customer churn prediction," pp. 1-4, 2010.

[7] K. C. a. D. V. d. Poel, "Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques," vol. 34, pp. 313-327, 2008.

[8] A. S. a. P. K. Panigrahi, "A neural network based approach for predicting customer churn in cellular network services," vol. 27, pp. 26-31, 2011.

[9] Y. Y. L. C. a. S. Z. X. Hu, "Research on a customer churn combination prediction model based on decision tree and neural network," pp. 129-132, 2020.

[10] S. B. a. S. Srivatsa, "Naive bayes classification approach for mining life insurance databases for effective prediction of customer p over life insurance products," vol. 51.

[11] L. F. a. H. Wang, "Estimating insurance attrition using survival analysis", Casualty Actuarial Society," vol. 8, pp. 55-72, 2014.

[12] D. V. d. P. a. B. Lariviere, "Customer attrition analysis for financial services using proportional hazard models," vol. 157, pp. 196-217.

ISSN (online): 2581-3048
Volume 7, Issue 10, pp 333-343, October-2023
https://doi.org/10.47001/IRJIET/2023.710045

[13] T. L. O. Goonetilleke, "Mining life insurance data for customer attrition analysis," vol. 1, pp. 52-58.

[14] K. C. a. Y. X. W. H Au, "A novel evolutionary data mining algorithms with applications to churn prediction," pp. 532-545.

[15] V. S. a. D. K. Satpathi, "An Exploratory Study on the Use of Machine Learning Techniques in Insurance Industry".

[16] G. D. O. O. a. S. Cai, "A hybrid churn prediction model in mobile telecommunication industry," vol. 4, pp. 55-62, 2014.

[17] G. M. F. M. Qazi, "Predictive Modeling in Insurance: A Survey," vol. 50, pp. 68:1-68:36, 2018.

[18] R. D. O. H. R. O. a. H. F. Amjad Hudaib, "Hybrid data mining models for predicting customer churn," vol. 8, pp. 91-96, 2015.

[19] Y. X. X. L. a. W. Y. EWT Ngai, "Customer churn prediction using improved balanced random forests," vol. 36, pp. 5445-5449, 2008.

[20] L. X. X. Y. a. J. E. Ying Weiyun, "Preventing customer churn by using random forests modeling," pp. 429-434, 2008.

[21] M. B. a. S. Krummaker, "Prediction of claims in export credit finance: a comparison of four machine learning techniques," vol. 8, 2020.

[22] L. I. a. S. Zeadally, "Healthcare Insurance Frauds: Taxonomy and Blockchain-Based Detection Framework (Block-HI)," vol. 23, pp. 36-43, 2021.

[23] I. Matloob, "Sequence Mining and Prediction- Based Healthcare Fraud Detection Methodology," vol. 8, pp. 143256-143273, 2020.

[24] E. Alamir, "Motor Insurance Claim Status Prediction using Machine Learning Techniques".

[25] A. V. C. Team, "Applications of Machine Learning and AI in Insurance".

[26] S. I. Inc, "Predictive Modeling with Imbalanced Data: An Application to Bank Telemarketing Response," 2018.

[27] V. Ganganwar, "An Overview of Classification Algorithms for Imbalanced Datasets".

[28] K. L. L. C.-Y. J. Peng, "An Introduction to Logistic Regression Analysis and Reporting," vol. . 96, pp. 12-23, 2002.

[29] A. Usman, "Binary Logistic Regression Analysis on Admitting Students using JAMB Score".

[30] T. R. a. C. Lo, "Determinants of User Acceptance of Internet Banking: An Empirical Study".

[31] W. A. Kamakura, "Cross-Selling," *www.researchgate.net*, 2008.

[32] P. P. A. Anit, "Modeling complex longitudinal consumer behavior with Dynamic Bayesian networks: An Acquisition Pattern Analysis application," *ResearchGate*.

[33] G. M. F. M. Qazi, "Predictive modeling of insurance sales using multivariate adaptive regression splines," *ACM SIGKDD Explorations Newsletter*, vol. 18, pp. 22-32, 2016.

[34] K. A. B. M. L. H. "Use of statistical models to predict feed intake of beef cattle," vol. 89, pp. 3181-3191, 2011.

[35] "Introduction to Logistic Regression Analysis and Reporting," vol. 96, pp. 12- 16, 2002.

[36] A. Usman, "Binary logistic regression analysis on ADDMITING students using jamb score".

[37] K. L. L. C. J. Peng, "An Introduction to Logistic Regression Analysis and Reporting," vol. 96, p. 96, 2002.

[38] A. Usman, "Binary logistic regression analysis on ADDMITING students using jambscore".

AUTHORS BIOGRAPHY

Savindi Welikadaarachchi, a dedicated Data Analyst at Acuity Knowledge Partners Colombo Sri Lanka, is a data science enthusiast contributing to the world of IT with her strong knowledge in Python, SQL, Excel and Critical Thinking.

Mereesha Botheju, an intern data analyst and passionate about the field of Data Science, indicates a keen interest in working with data, analyzing patterns, and deriving valuable insights.

Dinushi Ariyasena, an intern Data Engineer, brings a wealth of knowledge and strong dedication to Python programming, data pipeline development, and deep learning.

Viruni Fernando, an Intern Data Analyst at Octopus BI, actively contributes to data science with a strong interest in Python, SQL programming and machine learning.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 10, pp 333-343, October-2023

https://doi.org/10.47001/IRJIET/2023.710045

Dr. Anjalie Gamage, a Senior Lecturer at the Sri Lanka Institute of Information Technology, is a highly accomplished academic professional with expertise in Computational Linguistics, AI, and E-Learning.

Miss. Poojani Gunathilake, Assistant lecturer at the Sri Lanka Institute of Information Technology, Sri Lanka.

Citation of this Article:

Savindi Welikadaarachchi, Mereesha Botheju, Dinushi Ariyasena, Viruni Fernando, Anjalie Gamage, Poojani Gunathilake, "Data Driven Approach to Improve Profitability in Vehicle Insurance Sector" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 7, Issue 10, pp 333-343, October 2023. Article DOI https://doi.org/10.47001/IRJIET/2023.710045

****

