Chapter 20.

CHALLENGES IN APPLYING MACHINE LEARNING FOR PREDICTIVE MODELLING

The rapid evolution of digital financial transactions and insurance operations has significantly increased the reliance on machine learning for predictive modelling. The application of sophisticated machine learning techniques, including feature transformation, data balancing, and model optimisation, enabled the detection of anomalies in financial systems and claim predictions in the insurance sector. Assuming that artificial intelligence (AI) can contribute to the improvement of the actuarial profession in the Republic of Srpska, this chapter of the monograph will, along with discussing its application in predicting claims and assessing insurance risk, also present the prerequisites that artificial intelligence needs to fulfil to be utilised in the insurance market of the Republic of Srpska while following ethical standards and actuarial practice guidelines. Our aim is to explore the potential impacts of artificial intelligence on the actuarial profession, analysing how actuaries can use AI tools and techniques to enhance their work and competencies and, thus, provide benefits to policyholders, insurers, and the development of this profession.

1. FEATURES OF THE INSURANCE MARKET OF THE REPUBLIC OF SRPSKA

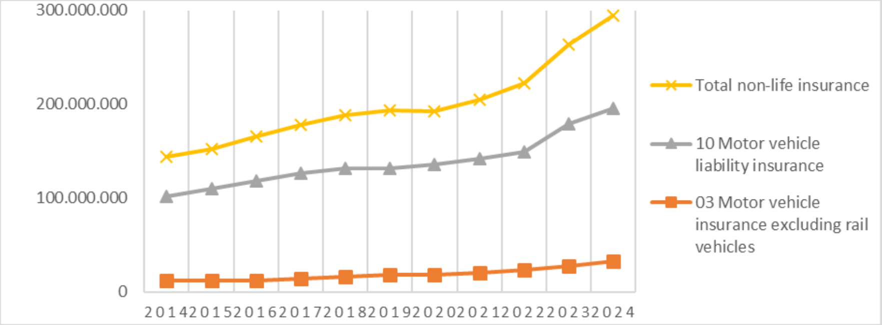
The insurance market in Bosnia and Herzegovina, like the state itself, is divided into two entities, each with its own supervisory body. In the Republic of Srpska, supervision of private insurance is performed by the Insurance Agency of the Republic of Srpska, established in 2006 under the provisions of the Law on Insurance Companies[578]. On the other hand, in the territory of the Federation of Bosnia and Herzegovina, the Insurance Supervision Agency of the Federation of Bosnia and Herzegovina has been operating since 2005 in accordance with the Law on Insurance Companies in Private Insurance[579]. Insurance companies register with one of the entities when establishing the company, and if they wish to expand their operations into the other entity, they must seek approval from the relevant supervisory authorities to conduct business activities in that entity.

[578] Law on Insurance Companies, *Official Gazette of the Republic of Srpska*, No. 17/05, 01/06, 64/06, 74/10, 47/17 and 58/19.

[579] Law on Insurance Companies in Private Insurance, *Official Gazette of the Federation of Bosnia and Herzegovina,* No. 24/05 and 36/10.

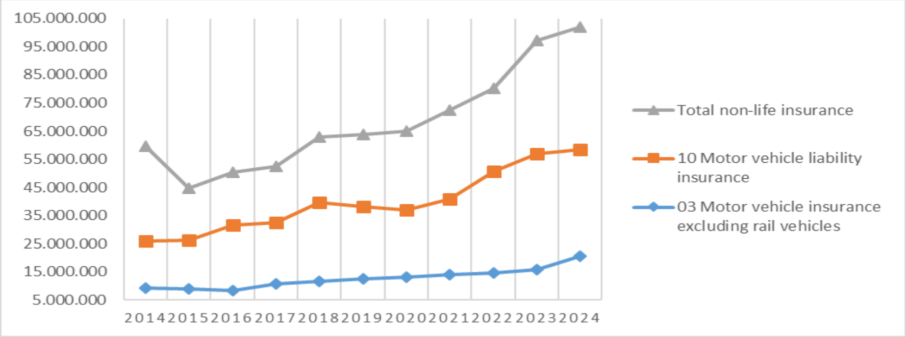
In 2024, 23 insurance companies operated in the insurance market of the Republic of Srpska, 14 of which were headquartered in the Republic of Srpska, while 9 were based in the Federation of Bosnia and Herzegovina and operated through their registered branches in the Republic of Srpska. In the Republic of Srpska insurance market, the total gross written premium for non-life insurance in 2024 amounted to BAM 294,583,748 (Figure 1), while the paid claims totalled BAM 102,075,429 (Figure 2). The car insurance premium (motor vehicle insurance excluding rail vehicles and motor vehicle liability insurance) accounted for approximately 77% of the total non-life insurance premium. These two types of insurance also accounted for 77% of the total paid claims in the Republic of Srpska market in 2024.

Figure 1. Non-life insurance premiums realised in the Republic of Srpska for the period 2014-2024



Source: Insurance Agency of the Republic of Srpska (2025)

Figure 2. Paid claims in the Republic of Srpska for the period 2014-2024



Source: Insurance Agency of the Republic of Srpska (2025)

Having presented the challenges in introducing innovative techniques in the Republic of Srpska insurance market, we will focus on motor vehicle liability Insurance and motor vehicle insurance, grouped under the term car insurance, in the example presented at the end of this section. The regulations and practices related to tariff setting in these two types of insurance currently differ in the market. Insurance companies in motor vehicle insurance or casco insurance set their own premium tariffs. Motor vehicle liability insurance belongs to the category of compulsory insurance regulated by the Law on Compulsory Traffic Insurance[580]. Article 12 of the law stipulates that the Agency Management Board shall implement a standardised premium tariff and price list for motor vehicle liability insurance, which remains in force until 31st December 2026. Along with the premium tariff, the insurance company is also obligated to submit to the Agency the technical bases it uses for setting premium rates. According to the applicable premium tariff and the price list for motor vehicle liability insurance in the Republic of Srpska when setting premium rates established by the tariff, the following criteria (risk factors) are taken into account: the type and purpose of a vehicle, technical characteristics of a vehicle (such as engine power, load capacity, engine displacement, number of registered seats, and number of employees in the workshop), duration of the insurance, and past performance of the insured (bonus/malus).[581] Following the full liberalisation of the motor vehicle liability insurance market, insurance companies in the Republic of Srpska will be able to incorporate additional factors into their tariff calculations, such as those listed in the section on Car Insurance Claim Classification. When considering which factors to include in tariff calculations, it is important to note that under the current Traffic Law, insurance contracts for third-party liability cover not only the vehicle owner but also any person using the vehicle with the owner's consent. Thus, the policy is vehicle-based rather than driver-based, and the motor vehicle liability insurance policy is not limited to a single driver. This practice is also practised by most European countries. On the other hand, in the United States, insurance policies are typically linked to a driver, which may explain why the focus is on driver-related factors when determining tariff models. We believe that current insurance regulations regarding the introduction of new products and amendments to the existing ones are necessary since liberalisation in setting prices could lead to destructive competition among insurers due to intense competition for market share and a decrease in premiums, threatening the solvency of insurance companies. Therefore, the liberalisation process should be handled carefully. Before introducing a new product or making amendments to

[580] Law on Compulsory Motor Vehicle Insurance, *Official Gazette of the Republic of Srpska*, No. 82/15, 78/20 and 1/24.

[581] Decision on the Standardised Premium Tariff and Price List for Motor Vehicle Liability Insurance in the Republic of Srpska, *Official Gazette of the Republic of Srpska,* No. 8/24, 91/24.

the existing one, insurance companies operating in the Republic of Srpska need to ensure the necessary criteria for effective risk management and control.[582]

In the following section, we will address the criteria for applying artificial intelligence (AI) concerning the future of the actuarial profession in the Republic of Srpska from the perspective of professional standards, the need for education in machine learning principles, and the provision of necessary data.

2. ACTUARIAL ASPECTS OF APPLICATION OF ARTIFICIAL INTELLIGENCE IN THE INSURANCE MARKET OF THE REPUBLIC OF SRPSKA

Considering the scope of the authorised actuaries' activities, their responsibilities, and the current regulations that govern them is vital in terms of analysing the assumptions, limitations, and risks related to the introduction of AI-based models in actuarial practice. Therefore, in this section, we will discuss some of the regulations that govern actuaries' practice in the Republic of Srpska. The Decision on the Content of the Opinion of a Certified Actuary[583] (Article 2) defines that the company is obliged to notify the Insurance Agency of the Republic of Srpska within 15 days that a new product has been adopted or has undergone amendments or supplements. Although the Insurance Agency of the Republic of Srpska does not give prior approval for the introduction of products pursuant to Article 13 of the Insurance Companies Act, it can limit the scope of insurance activities performed by a certain insurance company for a certain period if it is necessary to protect the financial capacity of the company. During the process of tariff control, the Insurance Agency used to require changes to the provisions related to discounts if they are not tied to actual and measurable risk characteristics that are the subject of insurance, and whose effects cannot be quantified or controlled in accordance with actuarial principles and methods. Article 54 of the Insurance Act defines that if the Insurance Agency of the Republic of Srpska determines that an insurance company is violating the rules of risk management and protection of policyholders, it can instruct the insurance company to change the types of insurance activities, suspend the application or amend the insurance terms and premiums, and take other measures necessary to improve risk management procedures.

[582] Mitrašević, M. (2016). Aktuarske odrednice razvoja proizvoda osiguranja. *Jahorina Business Forum, 2016*(1), pp. 445-459.

[583] Decision on the Content of the Opinion of the Certified Actuary, *Official Gazette of the Republic of Srpska*, No. 15/07.

The absence of regulations defining the principles for managing artificial intelligence and proper governance of operational risks related to digital security in the insurance market is currently, to some extent, compensated by the regulations in the areas of risk management, premium calculation, technical reserves, and capital adequacy. However, having recognised the potential of artificial intelligence for enhancing the actuarial profession in the Republic of Srpska and the risks associated with its application, outlined by Preez et al.[584] as well, regulating this area should become one of the key priorities in the Republic of Srpska insurance market.

Since there are no regulations directly defining the application of artificial intelligence in the Republic of Srpska, we will analyse the regulations in force in the European Union to conclude the potential risks posed by the lack of regulation in this area. In the European Union, the AI Act has been in force since July 2024, establishing a set of requirements that providers and users of high-risk AI systems are obliged to comply with. In the insurance sector, this law categorises AI systems used for risk assessment and pricing of individuals in life and health insurance as high-risk systems. The AI Act has introduced additional requirements for providers of high-risk AI systems.[585] The European Insurance and Occupational Pensions Authority (EIOPA) published a report in 2021 by the Consultative Expert Group on Digital Ethics, which outlines six principles for AI governance: proportionality, fairness and non-discrimination, transparency and explainability, human oversight, data management and record-keeping, and robustness and performance. Nevertheless, we will mostly focus on explainability, i.e. transparency, which implies the degree to which actuaries can understand and explain the decisions and predictions made by AI systems. Insurance companies should be able to explain to regulators and auditors the principles underlying their tariff models, and consumers should be informed about the key factors influencing the size of insurance premiums, enabling them to make informed decisions, adjust their choices, and accept the consequences. To achieve this, a high level of transparency and explainability is necessary for the systems, models, and data used. On the other hand, there may be fewer requirements for transparency in the case of fraud detection systems since fraud will always need to be proven. The goal of fraud detection systems is to provide

[584] du Preez, V., et al. (2024). From bias to black boxes: Understanding and managing the risks of AI – an actuarial perspective. *British Actuarial Journal*, *29*, e6. https://doi.org/10.1017/S1357321724000060

[585] European Parliament (2024). Artificial Intelligence Act: MEPs adopt landmark law. *Press release,* https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law

suggestions to the insurance company on how to prioritise certain actions and, accordingly, increase efficiency in internal processes.[586]

In addition to the lack of regulations governing the field of artificial intelligence in the Republic of Srpska, a significant limitation in conducting the research at the time of writing this section was caused by the lack of publicly available relevant data on experiences regarding artificial intelligence applications. The report on the integration of the digital economy in the Western Balkans shows that in 2021, 7% of businesses in this area used big data (3.9% in Bosnia and Herzegovina), 16% used cloud services (7% in Bosnia and Herzegovina), and 3% used artificial intelligence (2% in Bosnia and Herzegovina), compared to businesses in the European Union, where the percentages amounted to 14%, 34%, and 8%, respectively. Furthermore, based on publicly available data, we still cannot conclude the extent of the use of artificial intelligence in insurance companies in the Republic of Srpska.[587] According to our findings, experiences in applying artificial intelligence by certified actuaries in this market are mostly in form of ChatGPT as one of the large language models (LLMs). LLMs are a class of artificial intelligence models trained on large amounts of textual data and belong to the class of generative models. When using this tool, it is important to consider the research by Balona[588] that practically demonstrated that ChatGPT often provided incorrect mathematical facts and was unreliable with Python coding. A similar conclusion was drawn when GitHub's Copilot was used for coding assistance. Therefore, the need for careful review and potential optimisation of the generated code was emphasised. In accordance with professional standards, holders of actuarial functions in the Republic of Srpska should consider the potential impacts that large language models can have on business operations and conduct a careful review of the obtained answers and the decisions made based on them.

As for transparency, a different approach to the application of artificial intelligence in insurance for obtaining results was presented during the webinar

[586] European Insurance and Occupational Pensions Authority (2021). *Artificial intelligence governance principles: Towards ethical and trustworthy artificial intelligence in the European insurance sector*. Frankfurt am Main: EIOPA, https://www.eiopa.europa. eu/system/files/2021-06/eiopa-ai-governance-principles-june-2021.pdf

[587] Regional Cooperation Council (2023). *Western Balkans Digital Economy and Society Index (WB DESI) 2022 Report*, https://www.rcc.int/pubs/159/western-balkans-digital-economy-society-index-wb-desi-2022-report

[588] Balona, C. (2024). ActuaryGPT: Applications of large language models to insurance and actuarial work. *British Actuarial Journal, 29*, e15. https://doi.org/10.1017/S1357321724000102

titled Application of Artificial Intelligence in the Financial Sector of the Republic of Srpska, held on 25th March, 2025[589] on the premises of the Research Centre of the Faculty of Business Economics, University of East Sarajevo. Miona Graorac, Senior Actuarial Consultant in the company Willis Towers Watson[590], presented to the representatives of insurance companies a machine-led reserving algorithm developed by the company as part of its ResQ software. This approach is explained in more detail in the monograph published in 2024 by Mitrašević, Kočović, Koprivica, and Graorac.[591] Due to the fact that the software uses the methods that are widely accepted in actuarial practice in the Republic of Srpska, and since the model outputs are transparent, the use of this software is possible either without or with only minor amendments to the technical bases of insurance and the regulations governing the calculation of claim reserves in non-life insurance.

When applying artificial intelligence, insurance companies need to provide appropriate levels of supervision, adequate training for employees, define the tasks that need to be performed, as well as the persons responsible for performing AI-related tasks. It should be noted that the ability of insurance companies to adopt new technologies for insurance activities, pricing, and risk reduction depends on their ability to access sufficient, reliable, and high-quality external and/or internal data. According to our findings, a specific issue currently faced by domestic insurance companies is separate information systems for premium, financial accounting, and the information systems where claims are recorded; nevertheless, the link between those systems is not entirely automatised.[592] With the advancement of technology, insurers have been enabled to utilise new data sources. Wong et al. (2021) presented a review of the studies on telematics pricing as one of the emerging topics in actuarial science and practice, given that

[589] University of East Sarajevo, Faculty of Business Economics (2025). *Webinar held on the topic: Application of artificial intelligence in the financial sector of the Republic of Srpska*, http://www.fpe.ues.rs.ba/news/10043/257/odrzan-vebinar-na-temu-primjena-vjestacke-inteligencije-u-finansijskom-sektoru-republike-srpske.html

[590] Willis Towers Watson (n.d.). *Homepage*. https://www.wtwco.com

[591] Mitrašević, M, Kočović, J., Koprivica M., & Graorac, M. (2024). Application of artificial intelligence in projecting claims within non-life insurance. In: *Transformation of the Economy with Artificial Intelligence: Perspectives, Challenges and Opportunities*, Mitrašević, M. et al. (eds.), Bijeljina: University of East Sarajevo, Faculty of Business Economics, pp. 36-54.

[592] Mitrašević, M. (2019). Obezbeđenje kvaliteta podataka kao ključni preduslov adekvatne procene obaveza osiguranja, *EkonBiz*, *19*, pp. 292-302.

this type of data has only recently appeared.[593] The ability of insurance companies to apply new data sources, analytical tools based on artificial intelligence and machine learning is influenced by legislation, regulations, and supervisory measures implemented to ensure safeguards against unfair discrimination, and support digital security. It is also important to note that rigorous data protection protocols within insurance institutions can safeguard policyholders' privacy and preserve the integrity and reputation of the insurer.[594]

In accordance with generally accepted standards of practice, actuaries have to ensure that models are fit for their intended purpose and free from material biases, which in the case of machine learning models can arise from their training data or the selected algorithm. Understanding the model provides actuaries with essential tools to efficiently debug models and identify potential issues in data processing or model training. Furthermore, understanding the model is crucial for detecting biases that may have various impacts on certain groups. By uncovering such inappropriate biases, actuaries can take corrective measures, mitigate discrimination, and support fairness in decision-making processes. As for applying artificial intelligence for setting insurance prices, regulators would need to develop a test for checking the models of insurance companies to determine whether they comply with the relevant standards.[595]

3. EMPIRICAL IMPLEMENTATION OF MACHINE LEARNING IN INSURANCE

This study focuses on two key implementations: credit card fraud detection and car insurance claim prediction, leveraging extensive datasets to construct robust predictive models that mitigate financial risks and optimise operational efficiency.

[593] Blier-Wong, C., Cossette, H., Lamontagne, L., & Marceau, E. (2021). Machine learning in P&C insurance: A review for pricing and reserving. *Risks, 9*(4), 80. https://doi.org/10.3390/risks9010004

[594] Tešić, N., & Kočović De Santo, M. (2024). Opportunities for the application of artificial intelligence in managing catastrophic risks. In: *Transformation of the Economy with Artificial Intelligence: Perspectives, Challenges and Opportunities,* Mitrašević, M. et al. (eds.), Bijeljina: University of East Sarajevo, Faculty of Business Economics.

[595] Paunović, B., Tešić, N., & Kočović, J. (2019). Impact of Industrial Revolution 4.0 on insurance and its contribution to sustainable development. In: *Contemporary trends in insurance at the beginning of the fourth industrial revolution*, Kočović, J. et al. (eds.), Belgrade: University of Belgrade, Faculty of Economics and Business, pp. 3-19.

Credit Card Fraud Detection

In this section, we will demonstrate the possibilities of using machine learning in credit card fraud detection, as one of the highly significant methods for managing the risk of insuring financial losses due to credit card data theft. In 2025, Hafez, Hafez, Saleh, et al. published a systematic review of the studies included in the Scopus database that dealt with credit card fraud detection. Their review showed that this topic was covered in as many as 628 studies, but they based their analysis on 52 studies according to the inclusion and exclusion criteria.[596] Credit card fraud remains a rising concern in the U.S. financial landscape, with recent data specifying that 63% of credit card holders have been victims of fraud. In 2023, over 62 million Americans reported unauthorised charges, resulting in losses exceeding $6.2 billion annually. Furthermore, only 8% of these cases involved physically stolen or lost cards, with the majority arising from remote access to account credentials and sensitive personal data.[597] Traditional fraud detection methods, reliant on predefined rule-based approaches, often suffer from inefficiencies, producing high false-positive and false-negative rates. To address these limitations, machine learning-based solutions have been developed to enhance fraud detection accuracy by learning complex transaction patterns and identifying anomalies in real time.[598]

The dataset utilised in this study was sourced from the Kaggle dataset website and comprises 284,807 transactions across 31 features.[599] A key characteristic of this dataset is the application of Principal Component Analysis (PCA) to transform 28 of these features (V1-V28). The decision to employ PCA stems from the need to anonymise sensitive transactional data while preserving the statistical integrity of the dataset. PCA effectively reduces dimensionality by capturing the most significant variance in the data, eliminating multicollinearity, and mitigating overfitting risks. Furthermore, PCA enhances computational efficiency, allowing models to process transactional data at scale while focusing

[596] Hafez, I. Y., Hafez, A. Y., Saleh, A., Abbas, A., Mostafa, A., & Abdelrahman, A. (2025). A systematic review of AI-enhanced techniques in credit card fraud detection. *Journal of Big Data, 12*, Article 6. https://doi.org/10.1186/s40537-024-01048-8

[597] Cruz, B. (2025). 62 million Americans experienced credit card fraud last year. *Security.org*. Retrieved February 17, 2025, from https://www.security.org/digital-safety/credit-card-fraud-report/

[598] Mohammed, M. A., Kothapalli, K. R. V., Mohammed, R., Pasam, P., Sachani, D. K., & Richardson, N. (2017). Machine learning-based real-time fraud detection in financial transactions. *Asian Accounting and Auditing Advancement, 8*, pp. 67-76.

[599] Kaggle (n.d.). *Homepage*. https://www.kaggle.com

on the most informative patterns.[600] The remaining three features in the dataset include Time, which represents the seconds elapsed between a given transaction and the first transaction in the dataset, Amount, which denotes the transaction value, and Class, which is the target variable where fraud is denoted by 1 and non-fraud by 0.

A fundamental challenge in fraud detection is the extreme imbalance in class distribution, with fraudulent transactions constituting only 0.17% of the dataset. Training machine learning models on such skewed data inherently biases predictions towards the majority class, leading to poor fraud detection capabilities. To counteract this, Synthetic Minority Over-Sampling Technique (SMOTE) and Random Under-Sampling (RUS) were employed. SMOTE addresses class imbalance by generating synthetic fraudulent transactions rather than duplicating existing ones. This is achieved by interpolating between minority class samples, thereby preserving feature relationships and increasing the diversity of fraudulent transaction representations within the dataset.[601] By contrast, RUS mitigates class imbalance by randomly removing instances from the majority class, reducing computational complexity but at the risk of discarding valuable information. A balanced dataset was ultimately achieved, consisting of 4,920 non-fraudulent transactions and 2,460 fraudulent transactions.[602]

Exploratory data analysis revealed distinct behavioural differences between fraudulent and non-fraudulent transactions. On average, fraudulent transactions occur in shorter time intervals compared to legitimate ones, indicating a pattern of rapid, unauthorised purchases. Additionally, the mean transaction amount for fraudulent cases is significantly higher ($122.2) than for non-fraudulent cases ($88.3), reinforcing the notion that fraudsters typically execute high-value transactions before detection mechanisms can intervene. To ensure the selection of the most predictive features, an Analysis of Variance (ANOVA) test was conducted, ranking features based on their statistical relevance to the target variable. Features with an ANOVA score below 50 were discarded, allowing the model to focus on attributes that exhibit the strongest discriminatory power between fraudulent and non-fraudulent transactions.

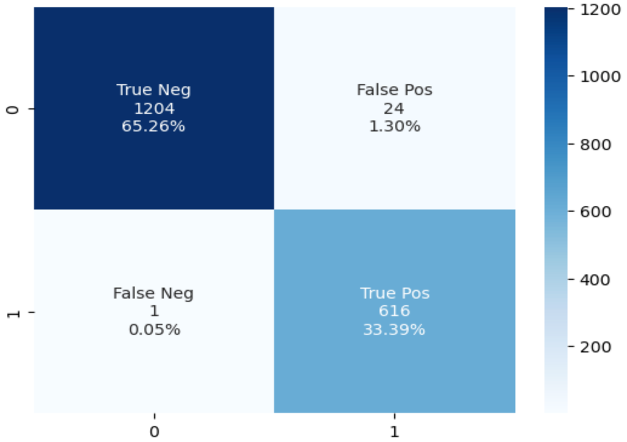
[600] Jolliffe, I. T. (2002). *Principal component analysis,* 2nd ed., Springer.

[601] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16,* pp. 321-357.

[602] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), pp. 1263-1284.

The dataset was subsequently partitioned into training and testing sets, with a 75%-25% split. Given the synthetic nature of the minority class samples generated through SMOTE, model evaluation could not rely solely on accuracy, as this metric does not adequately reflect the model's ability to detect rare fraudulent cases. Instead, performance was assessed using Cross-Validation Score and the Receiver Operating Characteristic – Area Under the Curve (ROC-AUC) Score. Cross-validation was employed to ensure the model's generalizability across different data subsets by iteratively training and testing on multiple splits of the dataset. The model achieved a mean cross-validation accuracy of 98.37%, demonstrating its ability to maintain high predictive performance across different data partitions. The ROC-AUC score, which evaluates the model's trade-off between true positive and false positive rates, was measured at 0.9951, indicating excellent fraud detection capability. The model's false positive rate was recorded at 1.3%, meaning that 1.3% of legitimate transactions were incorrectly flagged as fraudulent. More critically, the false negative rate was 0.05%, ensuring that only a negligible percentage of fraudulent transactions went undetected.

Figure 3. Confusion Matrix



Source: Author's calculation

Table 1. ROC AUC score and cross-validation calculation

Mean cross-validation accuracy	0.9837
ROC AUC Score	0.9951

Source: Author's calculation

While the model successfully distinguishes fraudulent from non-fraudulent transactions, further improvements can be made by expanding the feature space to incorporate additional contextual variables. Governments and financial institutions should consider integrating transactional metadata such as geolocation data, timestamps, temporal gaps between successive transactions in different locations, client identification numbers, linked bank accounts, email addresses, and device information. The inclusion of transaction type classifications—such as cash withdrawals, e-commerce purchases, bill payments, and foreign transactions—would further refine fraud probability estimation. Beyond feature augmentation, the adoption of deep learning techniques, particularly Recurrent Neural Networks (RNNs) and Transformer-based architectures, could enhance fraud detection by capturing sequential dependencies in transactional behaviors. These advanced models excel in learning temporal patterns, enabling them to detect sophisticated fraud strategies that evolve over time. In addition to technological advancements, regulatory interventions are necessary to strengthen fraud prevention frameworks. Governments should enforce real-time transaction monitoring standards, mandate secure authentication mechanisms such as multi-factor authentication (MFA) and biometric verification, and establish data-sharing protocols among financial institutions to facilitate collaborative fraud detection efforts. The implementation of adaptive fraud detection algorithms that dynamically adjust risk thresholds based on transaction history and behavioural patterns would further enhance the robustness of fraud prevention systems. The model achieved a sophisticated level of predictive accuracy. Future research should focus on integrating richer transactional features, leveraging deep learning methodologies, and advocating for policy reforms to create a more secure and resilient financial ecosystem.

Car insurance claim classification

The ability to predict car insurance claims with high precision is an essential factor in optimising risk management strategies for insurance providers. Accurate claim prediction enables insurers to mitigate financial losses, adjust premium pricing models, and identify high-risk policyholders, thereby enhancing operational efficiency and customer satisfaction. Leveraging machine learning techniques for this task involves extensive data preprocessing, feature engineering, and model optimization to ensure robust predictive performance. This study implements three machine learning algorithms – Random Forest, Logistic Regression, and K-Nearest Neighbors (KNN), to construct a predictive framework using a dataset sourced from Kaggle, containing 10,000 observations and 32 features.

The dataset encapsulates a diverse set of policyholder attributes, including demographic characteristics such as AGE, GENDER, EDUCATION, and RACE, financial metrics such as INCOME and CREDIT_SCORE, as well as driving behaviour indicators such as SPEEDING_VIOLATIONS, DUIS, and PAST_ACCIDENTS. The target variable, OUTCOME, represents whether a policyholder has filed an insurance claim, thus formulating a binary classification problem. Initial exploratory data analysis revealed that claimants aged 26–29 with 0–9 years of driving experience constitute a significant proportion of insurance clients. Furthermore, claim behaviour varies across income brackets, with upper-class individuals primarily insuring sedans. The dataset exhibited missing values in CREDIT_SCORE and ANNUAL_MILEAGE, which were imputed using median values to preserve data integrity. Categorical variables (AGE, DRIVING_EXPERIENCE, EDUCATION, INCOME, POSTAL_CODE) were transformed using one-hot encoding to facilitate their inclusion in machine learning models. Given the inherent class imbalance in the dataset, where claim occurrences are underrepresented, SMOTE was employed to generate synthetic instances of the minority class, ensuring equitable representation in model training.

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of the classes for classification tasks. It is particularly advantageous for handling complex, non-linear relationships in large datasets and is robust to overfitting. However, its main drawback lies in its higher computational cost, and the model's interpretability can be limited due to the complexity of combining numerous decision trees.[603] The K-Nearest Neighbors (KNN) classifier works by assigning a new data point to the majority class of its k nearest neighbors based on distance metrics. This algorithm is advantageous for non-linear data distributions and does not require explicit training, making it a simple and intuitive model. However, KNN can become computationally expensive during the prediction phase, especially for large datasets, as it requires distance calculations for each new sample. Moreover, KNN's performance is highly dependent on the choice of the number of neighbors and the distance metric used.[604] Logistic Regression is a linear model that estimates the probability of a binary outcome by applying a logistic function to a set of features. This model is widely appreciated for its simplicity, interpretability, and the ability to provide statistical insights into how individual features contribute to the prediction. However, it assumes a linear relationship

[603] Breiman, L. (2001). Random forests. *Machine Learning*, *45*, pp. 5-32.

[604] Kataria, A., & Singh, M. D. (2013). A review of data classification using K-nearest neighbour algorithm. *International Journal of Emerging Technology and Advanced Engineering, 3*(6), p. 354.

between features and the outcome, which may limit its performance in more complex, non-linear scenarios.[605]

Table 2. Classification report of three different models

Model	Precision	Recall	F1-score	Support
Random Forest				
Class 0	0.8832	0.8724	0.8778	1317
Class 1	0.8838	0.8937	0.8887	1430
Accuracy			0.8835	2747
Macro Avg	0.8835	0.8831	0.8833	2747
Weighted Avg	0.8835	0.8835	0.8835	2747
K-Nearest Neighbors				
Class 0	0.8985	0.8064	0.8499	1317
Class 1	0.8371	0.9161	0.8748	1430
Accuracy			0.8635	2747
Macro Avg	0.8678	0.8612	0.8624	2747
Weighted Avg	0.8665	0.8635	0.8629	2747
Logistic Regression				
Class 0	0.8429	0.8231	0.8329	1317
Class 1	0.8405	0.8587	0.8495	1430
Accuracy			0.8416	2747
Macro Avg	0.8417	0.8409	0.8412	2747
Weighted Avg	0.8417	0.8416	0.8416	2747

Source: Author's calculation

To evaluate the effectiveness of the algorithms, several key metrics were considered. Precision measures the proportion of true positives among all instances classified as positive, making it essential for minimising false positives, important in scenarios such as insurance claims, where non-claimants should not be mistakenly classified as claimants. Recall, on the other hand, focuses on the proportion of actual positives that are correctly identified, ensuring that genuine claim cases are not overlooked. The F1-score, the harmonic mean of precision and recall, is particularly valuable when balancing both metrics is crucial, as it

[605] Zou, X., Hu, Y., Tian, Z., & Shen, K. (2019). Logistic regression model optimization and case analysis. In: *Proceedings of the IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, IEEE.

provides a unified measure of classification effectiveness.[606] While accuracy is often a commonly referenced metric, it can be misleading in imbalanced datasets. In these cases, the Area Under the ROC Curve (AUC-ROC) is an additional metric that was used to assess model performance, with Random Forest attaining the highest score of 0.89, confirming its superior classification power.[607]

From the results presented in Table 2, the Random Forest model exhibits the best overall performance. The precision for non-claimants (class 0) is 0.8832, and for claimants (class 1), it is 0.8838, with recall values of 0.8724 and 0.8937, respectively. The model's F1-scores of 0.8778 and 0.8887 further highlight its ability to balance both metrics effectively, confirming its status as the top-performing model. The KNN classifier demonstrated a strong recall performance of 91.61% for claimants (class 1), showing its ability to identify claim-prone policyholders. However, the precision for class 1 is lower at 83.71%, which indicates that the model is prone to classifying non-claimants as claimants, resulting in false positives. This is reflected in its overall F1-score of 0.8624, which indicates a solid balance between recall and precision, but with a tendency to misclassify non-claimants as claimants. The accuracy of 86.35% underscores its strength in identifying claimants, but it comes at the expense of precision for non-claimants. Logistic Regression, serving as the baseline model, showed slightly lower performance than Random Forest, with an accuracy of 84.16%. The precision for non-claimants is 0.8429, and for claimants, it is 0.8405, while recall is 0.8231 and 0.8587, respectively. Although its simpler structure offers better interpretability and insight into how individual features contribute to the prediction, it was less effective in accurately classifying both claimants and non-claimants compared to the more complex models. In summary, Random Forest offers the best overall performance, particularly in balancing precision and recall. Interestingly, when Hanafy, M., & Ming, R. (2022) applied ML methods: logistic regression, XGBoost, random forest, decision trees, naïve Bayes, and K-NN to predict claim occurrence on the dataset given by Porto Seguro, a large Brazilian automotive company, they also concluded that Random Forest proved better compared to the other methods.[608] KNN excels in recall but suffers from a trade-

[606] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, *10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

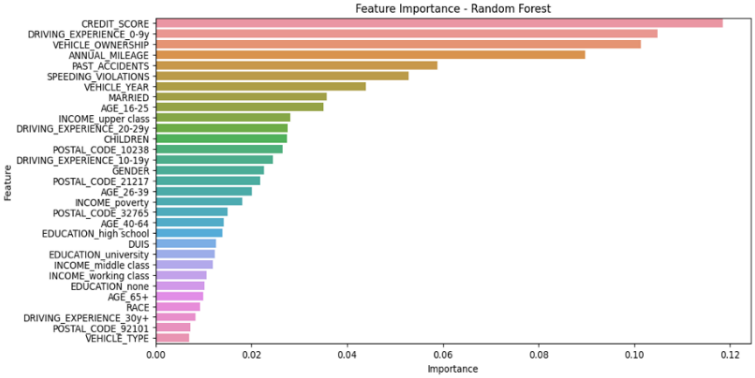
[607] Fernández-Delgado, M., Cernadas, E., Barro, S., & Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, *15*(1), pp. 3133-3181.

[608] Hanafy, M., & Ming, R. (2022). Classification of the insured using integrated machine learning algorithms: A comparative study. *Applied Artificial Intelligence*, *36*(1), pp. 1-18. https://doi.org/10.1080/08839514.2021.2020489

off in precision, particularly for non-claimants, while Logistic Regression, though interpretable, performs slightly worse than the more complex models in terms of both accuracy and classification effectiveness.

Regarding Figure 4, feature importance analysis demonstrated that CREDIT_SCORE, DRIVING_EXPERIENCE_0-9y, VEHICLE_OWNERSHIP, and ANNUAL_MILEAGE had the highest predictive power.

Figure 4. Feature Importance in Random Forest model



Source: Author's calculation

While the current models yield substantial predictive accuracy, incorporating external data sources and advanced modelling techniques could further enhance claim prediction reliability. Telematics data integration through real-time vehicle sensors, including braking patterns, acceleration behaviour, and mileage tracking, would enable a more dynamic assessment of risk profiles. The use of GPS tracking data would allow insurers to analyse driving patterns, road conditions, and accident-prone zones, correlating claim likelihood with environmental risk factors such as urban congestion, hazardous weather conditions, and high-speed corridors. Additionally, incorporating historical claim patterns, policy renewal history, and social determinants such as employment sector and travel frequency could refine risk stratification.[609] Deep learning architectures, including recurrent neural networks (RNNs) and transformer-based models, offer another avenue for improvement by capturing sequential dependencies in driving behaviour, financial transactions, and claim history. Autoencoders could enhance anomaly detection, aiding in the identification of fraudulent claims. Furthermore, regulatory and behavioural economics insights should be incorporated into predictive models to account for the impact of policyholder psychology, incentive

[609] Baecke, P., & Bocca, L. (2017). The value of vehicle telematics data in insurance risk selection processes. *Decision Support Systems*, *98*, pp. 69-79.

structures, and compliance frameworks. Reinforcement learning techniques may be applied to optimise dynamic pricing models based on real-time behavioural adjustments.[610] The empirical findings of this study underscore the potential of machine learning in optimising insurance risk assessment. By integrating multimodal data sources and leveraging advanced algorithmic techniques, insurers can achieve a more granular and dynamic understanding of claim probabilities, ultimately leading to more efficient claims processing, reduced fraudulent activities, and personalised policy offerings.

The key role of actuaries is to ensure that advanced modelling techniques are appropriately tailored to meet the criteria necessary for their sound and controlled use, to understand and manage the outcomes of these models, while acting in the public interest to ensure that such techniques are applied ethically and responsibly.

The source codes and implementation details for the machine learning models used in this study can be accessed via the following GitHub repositories:
https://github.com/bradickristina/frauddetection.git
https://github.com/bradickristina/CarInsuranceClaims.git

ACKNOWLEDGEMENT

This research is supported by the Ministry of Scientific and Technological Development and Higher Education of the Republika Srpska under the Agreement on Co-financing of the Scientific and Research Project, No. 19.032/961-46/24 of 30. December 2024.

[610] Boylan, J., Meyer, D., & Chen, W. S. (2024). A systematic review of the use of in-vehicle telematics in monitoring driving behaviours. *Accident Analysis & Prevention, 199*, 107519.