National College of Ireland

Vehicle Insurance Claim frequency and Amount Prediction through Machine Learning and Vehicle Analytics

MSc Research Project
Data Analytics

Arun Gangaramrao
Student ID: x22169202

School of Computing
National College of Ireland

Supervisor: Arghir Nicolae Moldovan

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Arun Gangaramrao
Student ID:	x22169202
Programme:	Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Arghir Nicolae Moldovan
Submission Due Date:	14/12/2023
Project Title:	Vehicle Insurance Claim frequency and Amount Prediction through Machine Learning and Vehicle Analytics
Word Count:	7400
Page Count:	22

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Arun Gangaramrao
Date:	31st January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Vehicle Insurance Claim frequency and Amount Prediction through Machine Learning and Vehicle Analytics

Arun Gangaramrao
x22169202

Abstract

This study explores the application of machine learning models for forecasting auto insurance claim severity and amounts across diverse datasets, optimizing risk assessment and claim processing. Leveraging three datasets, classification algorithms such as Logistic Regression, Random Forest, and Adaboost, and regression algorithms including Gradient Boost, Random Forest, SVR, Decision Tree, and Bayesian Regression are employed. While Adaboost faces challenges, Logistic Regression and Random Forest excel in multiclass classification and handle imbalanced classes well. In binary classification tasks, Random Forest consistently demonstrates superior performance across all datasets, achieving an impressive average accuracy of 98. On the other hand, when predicting claim amounts in regression tasks, Decision Tree emerges as the standout performer, particularly excelling in dataset 2 with a remarkable Mean Absolute Error (MAE) score of 77.49. Notably, Random Forest Regressor exhibits exceptional results in dataset 1, surpassing other models in accuracy and prediction effectiveness. The findings underscore the importance of considering dataset-specific features and class imbalances in model selection, providing valuable insights for improving predictive capabilities. Future work is proposed to further enhance these applications through customized model extensions and a deeper understanding of class imbalances and dataset features.

Keywords Claim Frequency, Usage Based Insurance, Machine Learning, Claim Amount, Classification, Regression

1 Introduction

1.1 Background

Increased vehicles on the road stem from the automotive industry's complexity and tech advancements. This has sparked heightened interest in auto insurance, driven by the need to manage risks like accidents, theft, vehicle damage, bodily injuries, and liability for damages by other parties Dewi et al. (2019).

The exact modeling of driving behavior has been made easier by developments in Deep Learning (DL) and the greater availability of linked vehicle data. Risk profiling for drivers, especially aggressive driving and context-related hazards, has the potential to lower accident rates and have a good social impact McDonnell et al. (2023).

The Insurance Information Institute reports an upward trend in both claim frequency and severity for US auto-insurances. Property damage claim severity increased by 11.5,

and frequency by 2.9 from Q1 2014 to Q1 2016. Moreover, average expenditures for US auto-insurance rose from USD 786.65 in 2009 to USD 889.01 by 2015 Fauzan and Murfi (2018).

1.2 Motivation

The motivation behind this research stems from the increasing complexity of risk assessment in the insurance sector. Advances in automotive technology, such as driver telematics, enable the insurance industry to incorporate new features into databases in addition to the ones that are already there. This coordinated effort improves the accuracy of risk assessments and claim forecasts, supporting a data-driven approach that has substantial advantages for both policyholders and insurers Peiris et al. (2023).

Traditional models fail to link driver behaviors and pricing models with the ultimate goal of maximizing company profits, especially under the real-world enterprise constraints He et al. (2018). An insurer's portfolio gains a temporal dimension through ongoing telematics data collection. This improves the appraisal of a driver's likelihood of filing a claim as well as their relative risk in the near future. When driving patterns from the past are successfully used to identify drivers who are at a higher risk of collision, swift actions to promote safer driving practices can help prevent accidents. Williams et al. (2022).

Predicting the frequency and amount of vehicle insurance claims is an important field of study with significant implications for the insurance sector. Understanding and forecasting claim trends is essential to promoting fair, open, and effective procedures as the insurance industry experiences revolutionary changes.

1.3 Research Question and Objectives

The primary research question driving this investigation is: **How can machine learning improve the accuracy of predicting vehicle insurance claim frequency and amounts using diverse datasets?**

In order to fill the gap, this study looks at a wider range of variables for more accurate risk assessments and insurance pricing. This work deviates from the established behavior-centric models by developing integrated models that incorporate both mobility-aware and demographic-aware components. In order to tackle this broad inquiry, the following particular research goals have been developed:

Smart Telematics Integration: By creatively integrating a sophisticated telematics integration, this project surpasses standard methods and offers a novel point of view that explores the complex aspects impacting both the frequency and amounts of claims in auto insurance. The incorporation of telematics is a revolutionary advancement in improving forecast precision.

Detailed Vehicle Insights: In contrast to standard studies, this research makes a notable contribution by extensively exploring specific vehicle features, transcending generic models. This innovative approach not only contributes to a more nuanced understanding but also adds a valuable dimension to the field, elevating the level of contribution in assessing how unique elements impact insurance claims.

Complete Risk Models: This work construct complete risk models and presents a novel approach by taking a wide range of parameters into account. By integrating factors like driving patterns, car attributes, and past insurance claims information, the comprehensive

approach contributes novelty and significantly advances the profession. The accuracy of predicting claim frequency is greatly improved by this thorough modeling method.

Next-Level Claim Prediction: Using fancy machine learning, our project predicts claim amounts considering factors like vehicle age, accident severity, and location. It's a step up from the usual, providing more precise estimations.

1.4 Structure of the Report:

The remainder of this report is structured as follows: Section 2 provides related work, delving into existing models and methodologies for insurance claim prediction. Section 3 outlines the data collection and preprocessing methodologies. In Section 4, Design Specification of the machine learning models for claim frequency prediction are detailed, followed by the implementation of the proposed solution in Section 5. Section 6 presents the evaluation metrics and results. Finally, Section 7 concludes the report, summarizing key insights and suggesting avenues for future research.

2 Related Work

Name	Algorithms Used	Key Findings and Results
Dewi et al. (2019)	Random Forest (five folds)	99% accuracy when using all features. Using only 1/3 of the overall features still produced comparable accuracy.
Zhang (2021)	Linear Model, Random Forest, SVM, XGBoost, Neural Network, Gradient Boosting	XGBoost consistently outperforms GLMs on all data sets. Neural networks, deep learning, and random forests perform better than GLMs on data sets with more independent variables and strong variable correlation.
Poufinas et al. (2023)	SVM, Decision Trees, Random Forests, Boosting	Random Forest fed with the top-15 most relevant variables shows MAPE of 18.24. XGBoost following with a MAPE of 19.56.
McDonnell et al. (2023)	TabNet, GLMs, XGBoost	TabNet Precision 0.55, Recall 0.20, F1 score 0.30, AUC 0.86, Accuracy 0.97, M cof – 0.32. Matthews coefficient was introduced.
Williams et al. (2022)	XGBoost, Logistic Regression, Stacking Classifier	Mean AUROC - Logistic Regression: 0.69 $\pm$ 0.10, XGBoost: 0.70 $\pm$ 0.12, Stacking Classifier: 0.70 $\pm$ 0.11. SMOTE for class imbalance.
Peiris et al. (2023)	Poisson distribution and the canonical link function	Proposed solution better than Naïve, Traditional, Boosting, and Full models. PCA is used.
Huang and Meng (2020)	Bayesian nonparametric model	Proposed regression framework has MAE value 0.6680, RMSE 1.1369.

Table 1: Summary of Studies and its Results

2.1 Telematics-Based Approaches in Car Insurance Pricing:

Telematics-based methods redefine car insurance pricing strategies. In Yan et al. (2020) from CNN-HVSVM algorithm, convolutional neural networks analyze driving behavior to

classify policyholders into risk levels. Utilizing 10-fold cross-validation, the training set is divided for effective model training and testing. The research delves into applying Internet of Vehicles (IoV) technology to determine car insurance rates, incorporating convolutions, pooling, and nonlinear activation functions. The study emphasizes high-weighted risk factors, categorizing customers into five driving behavior risk levels: extremely low, low, medium, high, and extremely high.

In contrast, He et al. (2018) proposes the Profile-Price-Profit (PPP) approach, to predict insurance premiums, the authors utilize an insurance pricing model that consists of two components integrating mobility and demographic-aware components to optimize insurance premiums for different risk profiles. It is found that PPP provides nearly a 10 price decrease for low-risk drivers and a 43 increase for high-risk drivers compared to Pay-How-You-Drive (PHYD) pricing.

2.2 Machine Learning Models for Claim Frequency and Severity Prediction:

In Huang and Meng (2019), an evaluation of SVM, random forests, XGBoost, and neural networks is conducted, with XGBoost identified as the preferred model for risk classification. The study reveals that employing binned variables generally outperforms using the original variables. The proposed ratemaking framework not only achieves high prediction accuracy but also meets interpretability requirements for both regulators and insured individuals.

TabNet, a deep learning architecture, is presented in McDonnell et al. (2023) as a better model for insurance risk pricing and claims prediction, beating conventional models in terms of accuracy and interpretability. The efficiency of the models under varying preprocessing efforts is evaluated. As an innovative and successful insurance pricing model, TabNet stands out for its high accuracy in capturing the sparsity of claims data and its highly interpretable outcomes.

2.3 Evolution of Usage-Based Insurance Models:

The study by Liu et al. (2022) uses neural networks, SVM, k-NN, decision trees, naive Bayes, and neural networks for classification by dividing the human factor into two categories static human factor (traditional) and dynamic human factor (driver behavior). The Genetic Algorithm (GA) is utilized to optimize the parameters of the SVM model. Author employs the classification algorithms to establish an objective relationship between driving behaviors and driving risks.

The second study Cunha and Bravo (2022) uses Bagging GLM and traditional Generalized Linear Models (GLM) with Poisson distribution to predict claim frequency, demonstrating that incorporating telematics data, particularly the mileage variable, significantly improves the overall quality of both the classical GLM model and the Bagging-GLM models compared to using only traditional ratemaking variables.

In Bian et al. (2018), a behavior-centric pricing mechanism for commercial vehicle insurance is introduced, examining detailed driving behavior characteristics. The study notes a positive correlation between total premium and distance or duration driven. Ensemble learning techniques are employed, combining multiple classifiers to make predictions, enhancing the accuracy of risk assessment in commercial vehicle insurance.

Baecke and Bocca (2017) uses random forests, artificial neural networks, logistic regression, generalized linear models, and logistic regression to investigate the advantages of telematics data incorporation into motor insurance risk profiles. Together, the many experiments' use of unique pricing schemes, intricate behavioral variables, and a variety of algorithms advances UBI models.

2.4 Addressing Imbalanced Data and Model Interpretability:

Dewi et al. (2019) demonstrates the scalability of the Random Forest model in handling big data problems with fewer features, offering a solution for efficient data processing. The study also finds that using only 1/3 of the overall features can produce comparable accuracy to using all features, highlighting the scalability of the Random Forest model in handling big data problems.

Williams et al. (2022) assesses the accuracy of XGBoost, emphasizing the significance of hyperparameter tuning and the impact of imputation methods on model performance in claim prediction. The study utilizes oversampling techniques, specifically SMOTE (Synthetic Minority Oversampling Technique), to expand the decision boundary and aid in the recognition of the minority class. Also Fauzan and Murfi (2018) shows that XGBoost gives better accuracies in terms of normalized Gini than the other methods.

2.5 Advancements in Forecasting Motor Insurance Claims:

Poufinas et al. (2023) identify weather conditions and car sales as influential variables affecting claims, showcasing the importance of external factors. The author also finds that the registration of new cars was found to be one of the most significant predictors of insurance claims, as more new cars circulating led to an increase in accidents and total claims cost. The time lag and weather conditions, such as the lowest temperature, also impacted the claims expense.

Huang and Meng (2020) proposes a data integration technique, efficiently combining traditional and telematics data, improving the efficiency of insurance claims prediction.The paper explores a Gaussian mixture model based on Dirichlet process priors to predict insurance losses, addressing the limitations of traditional parametric models in describing the distribution of losses. An advanced updating algorithm of slice sampling is integrated to apply an improved approximation to the infinite mixture model, enhancing the accuracy of data fitting and extrapolating predictions.

2.6 Limitations and takeaway

The use of machine learning to predict insurance claims has shown promising results, particularly with telematics data boosting accuracy, and algorithms like XGBoost proving consistently effective. However, challenges include limited dataset diversity, making it hard to apply models broadly. Inconsistent evaluation methods and algorithmic biases also need addressing. Deep learning, like TabNet, brings innovation, but its interpretability and real-world application efficiency remain challenges. Overall, while machine learning holds great potential, addressing these issues is crucial for its successful integration into insurance claim prediction systems.

3 Methodology

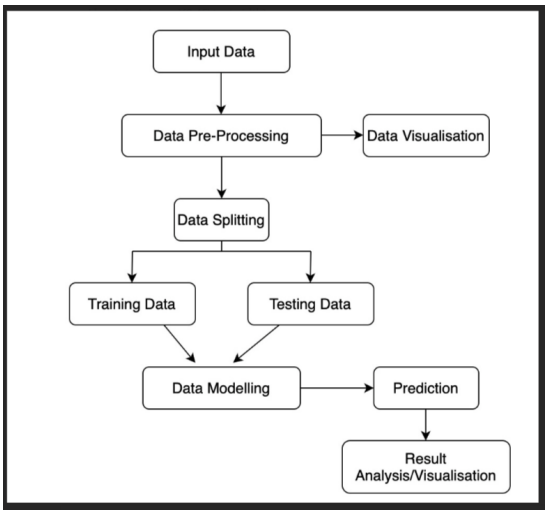

Figure 1: Flow chart representation of the project process

A methodical strategy is necessary to understand the complexity surrounding the frequency and quantities of insurance claims. This approach directs the process from data collection and preparation to the use of sophisticated machine learning. It describes procedures in great detail, including model selection, exploratory data analysis, and special sampling techniques. Figure 1 shows the Flow chart representation of the project.

3.1 Data Selection

Exploring the data and thoroughly understanding the domain are crucial steps before offering any solutions to real-world problems. For the study of the frequency of insurance claims and the claim amounts, a careful selection of data was carried out. By selecting different datasets from Kaggle[1], CAS[2] (Computational Actuarial Science using R) datasets, and synthetic data articles, this method aimed to increase the study's variety and the diversity.

Dataset 1: Synthetic telematics Data The dataset from So et al. (2021) includes 52 variables categorized into traditional features, telematics metrics, and response variables for claim count and amount. This synthetic telematics data that replicates main characteristics of a real telematics dataset so that the reproduced dataset can be shared with public without privacy concerns and proprietary issues Jeong (2022). For claim frequency prediction, only the claim count variable is utilized, while predicting claim amounts involves data only from claimed customers.

Dataset 2: French MTPL Dataset The second dataset merges data from 413,169 motor third-party liability policies, spanning a one-year duration. In freMTPLfreq, detailed risk features are paired with claim numbers, offering insights into policyholders' risk

[1] `https://www.kaggle.com/`
[2] `https://whttp://cas.uqam.ca/`

profiles. Complementing this, freMTPLsev intricately links claim amounts with corresponding policy IDs. FreMTPLfreq is exclusively utilized for predicting claim frequency, while the combination of freMTPLfreq and freMTPLsev is employed for predicting claim amounts Noll et al. (2018).

Dataset 3: Car insurance claim prediction dataset The third dataset unfolds with 44 columns and 58,592 rows, revealing insights into policyholder attributes. This expansive dataset includes crucial details like policy tenure, car age, owner's age, city population density, car make and model, and engine specifications. Notably, it features a binary target variable signifying whether a policyholder files a claim in the following six months, making it a vital resource for understanding claim filing behavior.

3.2 Data Preparation for Classification:

Gaining a comprehensive knowledge of the complex structures and patterns hidden in the information is our main goal using a wide range of datasets, including real-world insurance datasets from Kaggle and CAS as well as synthetic data. In this project for finding the frequency of insurance claims exploratory data analysis involved visualizing the class distribution using a countplot to understand the balance between claim and non-claim instances as shown in Figure 2Figure 3Figure 4.

Histograms were plotted for numerical features to gain insights into their distributions. For finding the insurance claim amounts the EDA invloved visualizing the correlation matrix heatmap is used to visualize their relationships with the dependent variable in a single plot.

Data Information and Missing Values: In order to create a foundational understanding, fundamental statistics and important details about the dataset were investigated. Analysis of distinct values related to the target variable was done. To guarantee data completeness, a thorough identification of missing values was also carried out.

Encoding Categorical Variables: In order to ensure compatibility with machine learning models, label encoding was developed as a crucial preprocessing step to transform categorical variables into a numerical representation.

Correlation Analysis: As shown in Figure 5Figure 6Figure 7A bar plot was used to investigate the correlation between the independent variables and the target variable, giving a graphic representation of their relationships.

Feature Selection: Variance threshold values of 0.1 were used to pick features based on their variance, with a focus on characteristics with significant variability.

Normalization: Standard scalar [3] normalization was applied to the dataset. This technique transforms the distribution of data, ensuring a standardized scale with a mean of 0 and a standard deviation of 1.

[3]Standard Scalar: `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html`

Sampling Technique Insurance claims prediction is an example of a classification task in which data are almost always strongly class imbalanced. Class imbalance may lead machine learning algorithms to exhibit a bias toward the dominant class, impeding effective learning of minority classes Haixiang et al. (2017). A dataset is considered to be class imbalanced if one class occurs much more often than the others. In a binary classification task the minority class is often referred to as the positive class, while the majority class is called the negative classWilliams et al. (2022).

To counteract this issue, SMOTE [4] (Synthetic Minority Over-sampling) technique, is employed. These strategies are instrumental in mitigating the impact of imbalanced data, fostering a more nuanced and accurate modeling outcome.

Model Selection: Model selection is a crucial process in machine learning that involves determining which algorithms are best suited for a certain prediction task. The type of data, the complexity of the relationships within it, and the specifications of the problem all have an impact on the models that are selected. Three different classification models were selected in this instance: the Adaboost classifier, the Random Forest classifier, and the Logistic Regression classifier.

Model Training and Evaluation: Train and Test Split [5] this partitioning, implemented using the Sklearn package, follows a 80:20 ratio. Following the model selection stage, the processed dataset is used to train the selected models. In order to teach the models the underlying patterns and correlations, historical data is presented to them. A variety of performance metrics are used to assess the models after the training phase, including as accuracy, AUC value and F1 Score. Receiver Operating Characteristic (ROC) curves and confusion matrices are also produced. The best model for predicting insurance claims can be chosen with the help of these visualizations, which offer a thorough grasp of how well the models distinguish between claim and non-claim occurrences.

3.3 Data Preparation for Regression:

Data Loading and Merging For the synthetic telematics data, Only the NBClaim column which is equal to 1 is selected from the dataset for predicting the insurance claim amounts. For the French MTPL data the code begins by loading two datasets, freMTPLfreq.csv and freMTPLsev.csv, using the Pandas library. These datasets are merged using the 'PolicyID' column as the common identifier, creating a consolidated DataFrame with information from both datasets.

Handling Missing Values A check for missing values in the Target variable column is performed. Any missing values in this column are then filled with zeros. This step is crucial for ensuring completeness in the dataset and avoiding complications during the modeling process.

[4]SMOTE: `https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html`

[5]TrainTestSplit: `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html`

Data Exploration and Encoding To prepare categorical data for machine learning models, numerical and categorical columns are identified within the datasets. The categorical columns is encoded using the LabelEncoder [6] from Scikit-learn, creating new columns and deleting the previous ones. This step facilitates the conversion of categorical values into a format suitable for regression models.

One-Hot Encoding [7] For the French MPTL data the categorical columns 'Brand', 'Gas', and 'Region' undergo one-hot encoding, transforming them into binary columns to represent different categories. The OneHotEncoder from Scikit-learn is employed for this purpose. The resulting one-hot encoded columns are then concatenated with the original dataset.

Correlation Matrix Heatmap To understand the relationships between different variables in the dataset, a correlation matrix heatmap is generated using Seaborn. This visualization provides insights into how various features correlate with each other. The size and color of each heatmap cell indicate the strength and direction of the correlation.

Custom Error Metrics Functions The code defines two custom error metric functions: Relative Absolute Error (RAE) and Root Relative Squared Error (RRSE). These functions will later be used to evaluate the performance of regression models in a way that is specific to the insurance claim prediction task.

Model Training and Evaluation Train and Test Split, this partitioning, implemented using the Sklearn package, follows a 80:20 ratio. Gradient Boosting Regressor, RandomForestRegressor, Support Vector Regressor (SVR), DecisionTreeRegressor, and BayesianRidge are considered for the regression task. GridSearchCV [8] is employed for hyperparameter tuning, optimizing the models for better predictive performance.

The evaluation metrics for each model, including Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R-squared (R2) Score, Relative Absolute Error (RAE), and Root Relative Squared Error (RRSE), are calculated and printed.

4 Design Specification

The predictive modeling for insurance claim frequency and claim amount involves a comprehensive design strategy that encompasses both classification and regression tasks. This design specification outlines the selected models, evaluation metrics, and other pertinent considerations for the classification and regression aspects of the insurance claim prediction.

[6]LabelEncoder: `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html`

[7]One-hot Encoder: `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html`

[8]GridSearchCV: `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html`

4.1 Modelling Technique

Logistic Regression: Logistic Regression [9] serves as a fundamental classification model in predicting insurance claim frequency. Employing the sigmoid function, it estimates the probability of a claim occurrence.As mentioned in Arunkumar and Yellampalli (2017) It is a technique for probabilistic view of classification. It is used to compute the possibility of a dichotomous outcome based on one or more independent variables which are called as predictor or features.

Random Forest: As mentioned in Dewi et al. (2019)The Random Forest model is an ensemble model,which is a machine learning technique that aims to combine predictions from several basic estimators to improve the accuracy of some of these basic estimators. Random Forest[10] is an ensemble model that uses a bagging approach, which is to build several basic models independently and the final prediction is obtained by looking for the average value (for regression cases) or voting (for classification cases) of each of these basic models Biau (2012).

AdaBoost: Fauzan and Murfi (2018) AdaBoost, an adaptive boosting technique, will be employed to sequentially train models, emphasizing misclassified instances. This model will be trained on the same driving behavior features, iteratively improving its performance. AdaBoost's [11] ability to focus on challenging instances will be particularly crucial in enhancing the model's predictive capacity for insurance claim occurrences.

Gradient Boost Regression: As mentioned in Zhang (2021) The Boosting algorithm was first derived from the concepts of weak learning and strong learning. Its basic idea is to use a series of weak learners to fit the sample data. Each iteration is to fit the difference between the model prediction value and the data obtained in the previous iteration.Gradient Boost Regression [12], a powerful ensemble method, will be employed to predict insurance claim amounts. The model will be trained on telematics features related to driving behavior, and the target variable will represent the claim amount.

Support Vector Regression (SVR): Support Vector Machines [13](SVM) is a supervised machine learning algorithm that is used for both classification and regression tasks (Support Vector Regression–SVR). In classic regression, the main objective is to minimize the sum of the least squared errors Poufinas et al. (2023) Huang and Meng (2019).

Decision Tree Regression: Decision Trees [14] are a supervised machine learning algorithm that is used for both classification and regression tasks. It works by recursively

[9]Logistic Regression: `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html`

[10]Random Forest: `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`

[11]Adaboost: `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html`

[12]Gradient Boost Regressor: `https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html`

[13]SVR: `https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html`

[14]Decision Tree Regressor: `https://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeRegressor.html`

partitioning the data based on the most informative features Poufinas et al. (2023).

Bayesian Regression: Bayesian Regression [15], incorporating Bayesian principles, will be employed for predicting insurance claim amounts. Telematics features will be considered, and the target variable will represent the claim amount Huang and Meng (2020) .

4.2 Evaluation Techniques

The evaluation of models predicting insurance claim frequency involves employing several key metrics to ensure a comprehensive assessment of their performance. Accuracy serves as a primary indicator, reflecting the overall correctness of predictions and essential for evaluating the models' ability to accurately classify claim frequency. An important indicator that strikes a compromise between recall and precision is the F1 Score, which also offers insightful information about the trade-offs between false positives and false negatives. While the micro F1 score is used in binary classification for a comprehensive evaluation of overall performance, the weighted F1 score is applied in multiclass classification to resolve class imbalance. The ROC-AUC curve, representing the Receiver Operating Characteristic and the Area Under the Curve, offers a comprehensive view of the model's discriminatory ability between claim and non-claim instances.

Additionally, metrics like Relative Absolute Error (RAE), Mean Absolute Error (MAE), and R2 Value provide standardized measures of prediction accuracy and assess the overall fit of the models in predicting claim amounts. These metrics collectively ensure a robust and multifaceted evaluation of the models' effectiveness in predicting insurance claim frequency.

5 Implementation

A successful machine learning model needs to be developed by carefully planning and carrying out a number of steps. Every step of the development process needs to be carefully planned and carried out to guarantee the model's correct operation, smooth deployment, and effective use in real-world scenarios.

5.1 Tools Used

Python is one of the languages and tools utilized in this process, which makes use of well-known libraries including Pandas, Scikit-learn, Seaborn, Matplotlib, and Plotly. The Jupyter Notebook environment in which the project is organized offers an interactive and thoroughly described platform Alamir et al. (2021).

5.2 Data Processing

Data Inspection Upon initial examination, the dataset information is obtained using the .info() method, revealing three distinct data types: float64, int64, and object. Further insights are extracted using .describe(). To facilitate model training, categorical values undergo encoding via the Label Encoder.

[15]Bayesian Regression: `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.BayesianRidge.html`

Handling Missing Values Identification of missing and null values is conducted through .isna(), .isnull(), and .sum() functions, enabling a comprehensive overview of the dataset's data integrity.

Correlation Analysis and Feature Selection A correlation analysis is generated to examine relationships between target and independent variables. Employing a variance threshold of 0.1, essential features are selected for modeling, optimizing the dataset for predictive analytics.

Classification Dataset Refinement A targeted technique is utilized for binary classification in the setting of a classification dataset that has several unique values in the target variable. The dataset is streamlined for binary classification tasks by carefully removing target variable values greater than 1.

Normalization Using Standard Scalar The Standard Scaler is used to provide consistent feature scaling during normalization, a crucial preprocessing step that improves model performance and stability during training.

Regression Model Evaluation Metrics Customized error metric functions are introduced for regression tasks in order to evaluate model performance. These include the Root Relative Squared Error (RRSE) and Relative Absolute Error (RAE), which offer more detailed information about the precision and dependability of the regression model.

5.3 Hyperparameter Tuning

The process of improving a machine-learning model's specified parameter values prior to the training phase is known as hyperparameter tuning. The behavior and performance of the model are governed by these parameters. Carefully adjusting hyperparameters becomes crucial because of their impact on model results, especially in situations involving multiclass classification. The goal of the current research is to optimize performance and increase model efficacy by fine-tuning the following hyperparameters.

In the field of machine learning, many methods and algorithms are essential for resolving issues with multiclass classification and regression. When using Logistic Regression for multiclass scenarios, important parameters like "maxiter," which indicates the maximum number of iterations required for solver integration, and "solver," which indicates the optimization technique, are involved. The Random Forest Classifier uses the "nestimators" argument to specify how many trees are in the forest, and it uses the "randomstate" parameter as a seed to generate random numbers.

"Learningrate" for shrinkage, "maxdepth" for the maximum depth of individual trees, and "nestimators" for the number of boosting stages are crucial factors to take into account while using the Gradient Boosting Regressor. Similar to this, "nestimators," "maxdepth," and "minsamplessplit" parameters are used by the Random Forest Regressor to ensure the best possible forest formation. Factors such as "epsilon" in the epsilon-SVR model, "C" for regularization, and "kernel" for defining the kernel type are taken into account by the Support Vector Regressor (SVR).

To control tree depth and node splitting conditions, the Decision Tree Regressor parameters "maxdepth," "minsamplessplit," and "minsamplesleaf" are used. Finally, the regularization and optimization of the model are influenced by the parameters "alpha1,"

"alpha2," "lambda1," and "lambda2" that specify the form and scale parameters for the Gamma distribution prior in Bayesian Ridge Regression. Together, these characteristics enhance each algorithm's adaptability and efficiency in solving certain regression problems. Table 2 shows the values used for the respective hyperparameters used in the research.

Table 2: Hyperparameters for Classifiers and Regression Models

Classifier/Regressor	**Best Hyperparameters**
Logistic Regression	'random_state': 0, 'max_iter': 1000, 'solver': 'lbfgs', 'multi_class': 'multinomial'
RandomForest Classifier	'n_estimators': 100, 'random_state': 42
Gradient Boost Regressor	'learning_rate': 0.1, 'max_depth': 5, 'n_estimators': 200
Random Forest Regressor	'max_depth': None, 'min_samples_split': 2, 'n_estimators': 150
Support Vector Regressor	'C': 100, 'epsilon': 0.5, 'kernel': 'linear'
Decision Tree Regressor	'max_depth': 10, 'min_samples_leaf': 4, 'min_samples_split': 2
Bayesian Ridge	'alpha_1': 1e-07, 'alpha_2': 1e-07, 'lambda_1': 1e-05, 'lambda_2': 1e-07

6 Evaluation

The evaluation stage of the machine learning pipeline is critical because it determines the effectiveness of the model and verifies that it functions as planned. A crucial step in effectively assessing the model's performance is choosing relevant assessment metrics.

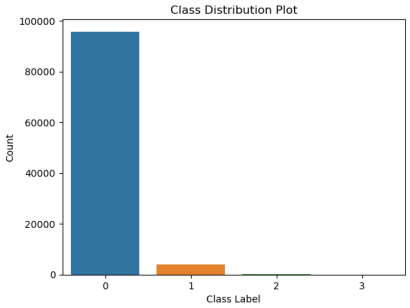

Figure 2: Synthetic telematics data class distribution

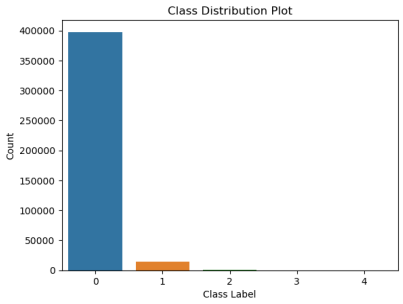

Figure 3: French MTPL data class distribution

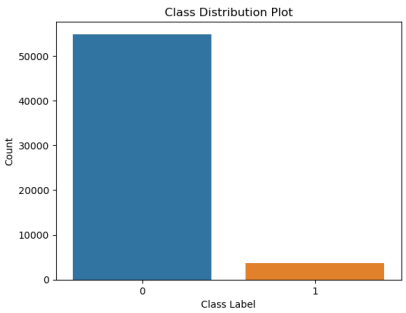

Figure 4: Car insurance dataset class distribution

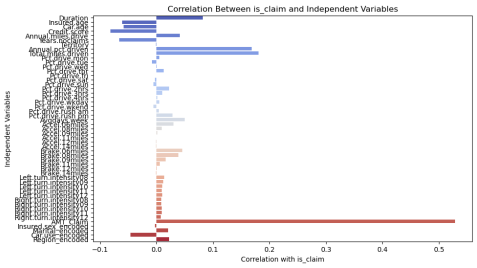

Figure 5: Synthetic telematics data Correlation Plot

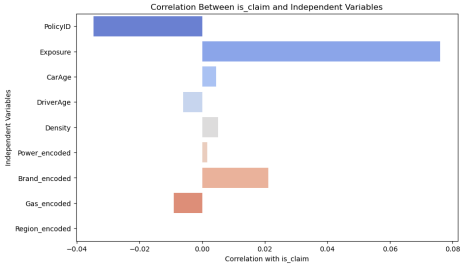

Figure 6: French MTPL data Correlation Plot

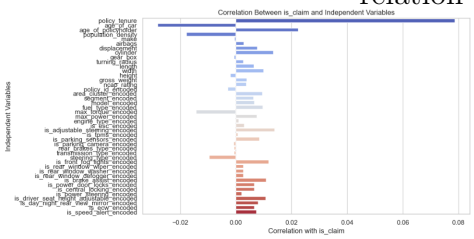

Figure 7: Car insurance dataset Correlation Plot

6.1 Dataset 1 insurance claim frequency prediction

In this case study, we evaluate the accuracy and the Area Under the Curve (AUC) metric of many classification models under diverse conditions. Both multiclass and binary classification problems are evaluated, taking sampling strategies and the existence or absence of class imbalance into account. The table 3 shows the values of the evaluation metrics carried out.

Table 3: Model Performance Metrics

Model	Accuracy (%)	AUC	F1 Score
MultiClass Without Sampling			
Logistic Regression	95.57	-	0.9352
Random Forest Classifier	96.90	-	0.9609
Adaboost Classifier	29.68	-	0.4438
MultiClass With Sampling			
Logistic Regression	69.67	-	0.6859
Random Forest Classifier	99.60	-	0.9960
Adaboost Classifier	50.11	-	0.4893
Binary Classification without Sampling			
Logistic Regression	95.82	0.75	0.9582
Random Forest Classifier	97.07	0.91	0.9707
Adaboost Classifier	95.87	0.80	0.9587
Binary Classification with Sampling			
Logistic Regression	71.49	0.77	0.7149
Random Forest Classifier	99.10	1.00	0.9910
Adaboost Classifier	83.21	0.92	0.8320

6.1.1 Multiclass Classification:

The Random Forest model is the best in multiclass classification, with the highest accuracy and F1 score, demonstrating its ability to handle challenging tasks. Remarkable generalization skills are displayed by the Random Forest Classifier, which effectively captures complex data relationships. With somewhat lower values without sample and even smaller values following sampling possibly because of oversampling issues that result in overfitting and decreased generalization, logistic regression closely resembles random forest. Conversely, the Adaboost Classifier trails behind, showing reduced accuracy and a worse F1 score, which may indicate a susceptibility to data noise.

6.1.2 Binary Classification:

Without sampling, Logistic Regression exhibits high discriminating ability for binary classification. The Random Forest Classifier uses its ensemble approach to determine feature relevance and performs exceptionally well, exhibiting excellent accuracy, F1 score, and AUC. In binary situations, the Adaboost Classifier exhibits adaptability and maintains its competitiveness. Sample variability causes problems for Logistic Regression when sampling is used, however it greatly helps the Random Forest Classifier, highlighting the significance of sampling in reducing class imbalance. The Adaboost Classifier further demonstrates the adaptability of its ensemble to unbalanced data by demonstrating improvement with sampling.

6.2 Dataset 2 insurance claim frequency prediction

As shown in the table 4A number of models demonstrate notable performances in the context of multi-class classification without sampling using the French Motor Third Party Liability (MTPL) data.

Table 4: Model Performance Metrics

Model	Accuracy (%)	AUC	F1 Score
MultiClass Without Sampling			
Logistic Regression	96.20	-	0.9433
Random Forest Classifier	96.09	-	0.9429
Adaboost Classifier	95.01	-	0.9374
MultiClass With Sampling			
Logistic Regression	48.59	-	0.4667
Random Forest Classifier	98.75	-	0.9874
Adaboost Classifier	43.92	-	0.4067
Binary Classification without Sampling			
Logistic Regression	96.45	0.63	0.9644
Random Forest Classifier	96.36	0.61	0.9635
Adaboost Classifier	96.45	0.66	0.9644
Binary Classification with Sampling			
Logistic Regression	48.59	0.63	0.5938
Random Forest Classifier	98.75	0.99	0.9697
Adaboost Classifier	43.92	0.78	0.7009

6.2.1 Multiclass Classification:

Logistic regression demonstrates its effectiveness in multiclass classification without sampling, as evidenced by its greatest accuracy of 96.20, which also highlights its adaptability to a variety of assignments. The Random Forest Classifier comes in second, excelling with an amazing F1 score of 0.9429 that shows it can successfully learn complex data associations. The scene changes, though, and the Random Forest Classifier emerges as the top performer with an incredible accuracy of 98.75 when sampling is applied. This highlights how flexible the Random Forest model is, particularly when dealing with imbalanced datasets and the requirement for reliable classification.

6.2.2 Binary Classification:

With an exceptional accuracy of 96.45 for binary classification without sampling, logistic regression stands out. With a marginally lower accuracy but an amazing F1 score of 0.9635, the Random Forest Classifier closely resembles Logistic Regression in the identical situation. The Random Forest Classifier performs quite well in the binary classification challenge with sampling, with an amazing F1 score of 0.9697 and a high accuracy of 98.75. All things considered, the Random Forest Classifier continuously exhibits strong performance in a variety of settings.

6.3 Dataset 3 Insurance Claim frequency prediction

6.3.1 Binary Classification without Sampling:

Considering table 5, when it comes to binary classification in the absence of sampling, Logistic Regression is a dependable option. Its accuracy of 93.55, balanced F1 score of 0.9354, and AUC of 0.60 demonstrate a pleasing combination of discrimination and accuracy in predicting claim frequency. With a respectable F1 score of 0.9309 and a strong accuracy of 93.10, the Random Forest Classifier has a marginally lower AUC of 0.57, indicating some difficulties in differentiating between positive and negative examples. With an accuracy of 93.55, a great F1 score of 0.9354, and an AUC of 0.63, Adaboost Classifier demonstrates its effectiveness as a reliable model for forecasting claim frequency without requiring sampling. In this scenario, Adaboost Classifier stands out for its balanced combination of accuracy, F1 score, and discrimination, closely followed by Logistic Regression.

6.3.2 Binary Classification with Sampling:

When sampling is added to binary classification, Logistic Regression shows a significant drop in accuracy to 57.99, a decent F1 score of 0.5799, and an AUC of 0.61. These results point to potential difficulties in managing imbalanced classes. In contrast, the Random Forest Classifier performs exceptionally well, exhibiting notable improvements with sampling and demonstrating its efficacy in handling imbalanced data with an amazing accuracy of 93.45, a strong F1 score of 0.9344, and an astounding AUC of 0.98. Adaboost Classifier exhibits a capacity to gain from sampling, although at a performance cost, as evidenced by its modest accuracy of 68.21, acceptable AUC of 0.76, and improved F1 score of 0.6820.

Table 5: Model Performance Metrics

Model	Accuracy (%)	AUC	F1 Score
Binary Classification without Sampling			
Logistic Regression	93.55	0.60	0.9354
Random Forest Classifier	93.10	0.57	0.9309
Adaboost Classifier	93.55	0.63	0.9354
Binary Classification with Sampling			
Logistic Regression	57.99	0.61	0.5799
Random Forest Classifier	93.45	0.98	0.9344
Adaboost Classifier	68.21	0.76	0.6820

6.4 Dataset 1 insurnace claim amount prediction

This case study delves into the assessment of regression models for predicting claim amounts. Five models, namely Gradient Boost, Random Forest, Support Vector Regressor (SVR), Decision Tree Regressor, and Bayesian Ridge Regression, are evaluated based on Mean Absolute Error (MAE), R-squared (R2), and Relative Absolute Error (RAE). Table 6 shows the values obtained during regression.

Table 6: Regression Model Performance Metrics

Metric	MAE	R2	RAE
Gradient Boost	1955.28	0.43	0.65613
Random Forest	1920.98	0.43	0.64462
SVR	2222.49	0.00	0.7458
Decision Tree Regressor	2409.33	-0.07	0.8085
Bayesian Ridge Regression	2475.86	0.12	0.83083

In the realm of regression modeling for predicting insurance claim amounts, the assessment of various algorithms reveals nuanced performances. Gradient Boost, achieving a balanced MAE, R2, and RAE, effectively captures intricate patterns in the data, making it a reliable choice. Random Forest, leveraging its ensemble approach, exhibits similar effectiveness but with slightly improved metrics, showcasing robustness in predicting claim amounts. However, the Support Vector Regressor (SVR) struggles, evident in its higher MAE and lackluster R2, indicating challenges in capturing the underlying data patterns. The Decision Tree Regressor fares poorly compared to ensemble methods, with a negative R2 suggesting it doesn't provide a meaningful improvement over a naive mean prediction. Bayesian Ridge Regression faces challenges in capturing data nuances, resulting in a comparatively higher MAE and RAE, reflecting limitations in accurately predicting insurance claim amounts. These insights guide the selection of regression models based on their performance nuances in specific contexts.

6.5 Dataset 2 insurnace claim amount prediction

Considering table 7In the evaluation of regression models for predicting insurance claim frequency, distinct performances emerge. Gradient Boost and Random Forest both exhibit suboptimal results with negative R2 values and high MAE, indicating challenges in capturing underlying data patterns and predicting claim frequency accurately. The Decision Tree Regressor shows some improvement with a lower MAE and slightly better

Table 7: Regression Model Performance Metrics

Metric	MAE	R2	RAE
Gradient Boost	77.49	-0.14	1.014
Random Forest	86.43	-1.53	1.014
Decision Tree Regressor	67.70	-0.06	0.88677
Bayesian Ridge Regression	106.42	0.06	1.393

R2, suggesting a more effective capture of certain patterns. However, the negative R2 still indicates limitations compared to a naive mean prediction. Bayesian Ridge Regression performs relatively poorly, facing challenges in capturing the complexities of claim frequency prediction, resulting in less accurate outcomes. These insights provide valuable considerations for selecting appropriate regression models in the context of predicting insurance claim frequency.

6.6 Discussion

Discussion of Insurance Claim Frequency Classification Experiments Predicting the frequency of insurance claims using three different datasets allowed for the performance of the classifiers to be better understood. Dataset 1 showed that Random Forest and Logistic Regression performed well in multiclass classification and were sensitive to unequal class distributions. Adaboost, however, has difficulty, particularly without sampling. All of the classifiers in Dataset 1 did well for binary classification, with Random Forest performing particularly well. These patterns were reflected in Dataset 2, which highlighted Random Forest's ongoing advantage against Adaboost. With an AUC of 0.98, Random Forest preserved robustness in Dataset 3, while Adaboost and Logistic Regression produced results that were similar. Class inequalities were a major problem that affected Adaboost, but Random Forest proved robust. Results from sampling procedures were inconsistent, indicating that using them should be done so with caution. Figure 8, Figure 9, Figure 10 show the confusion matrix and ROC curve of Random forest for the binary classification of three datasets.

The literature provides insightful information. For example, McDonnell et al. (2023) presents findings for Tabnet in Dataset 1 and shows remarkable accuracy of 97, F1 score of 0.30, and AUC of 0.86. Given Random Forest's superiority in binary classification and Tabnet's high accuracy, it may be an acceptable choice in some situations. Parallel to this, Williams et al. (2022) provides Mean AUROC results for Dataset 1, displaying the mean AUROC of 0.69 ($\pm$0.10) for Logistic Regression, 0.70 ($\pm$0.12) for XGBoost, and 0.70 ($\pm$0.11) for the Stacking Classifier. These results highlight the complexity of classifier performance even more, as the Stacking Classifier, XGBoost, and Logistic Regression all show competitive mean AUROC values.

Future research could benefit from investigating ensemble approaches, feature engineering, and sophisticated sampling strategies. A more thorough examination of class imbalances and their particular impact on various algorithms may provide information for customized model selection. The tests' findings highlighted the complexity involved in forecasting the frequency of insurance claims, highlighting the demand for advanced modeling techniques that take into account the particulars of each dataset and successfully resolve class imbalances.

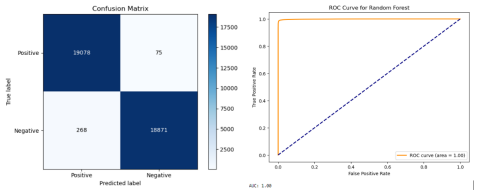

Figure 8: Synthetic telematics data Random Forest Confusion Matrix and ROC curve

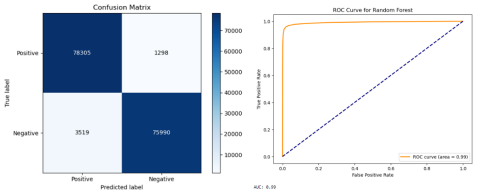

Figure 9: French MTPL data Random Forest Confusion Matrix and ROC curve

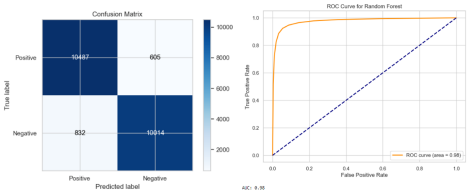

Figure 10: Car insurance data Random Forest Confusion Matrix and ROC curve

Discussion of Insurance Claim Amount Regression Experiments The outcomes of the tests conducted on the prediction of insurance claim amounts using two datasets offer fascinating new perspectives on how different regression models work. The Mean Absolute Error (MAE) values of the Gradient Boost and Random Forest in Dataset 1 were comparable, indicating equivalent accuracy. On the other hand, greater MAE values were shown by SVR, Decision Tree and Bayesian Ridge Regression, indicating less accuracy in claim amount prediction. For Bayesian Ridge Regression, Random Forest, and Gradient Boost, the R-squared (R2) values demonstrated positive correlation, suggesting a reasonable level of predictive power. Significantly, the Support Vector Regressor (SVR) showed an R2 value of 0.00, indicating that it was not predictive in this particular situation. The relative performance was supported by the Relative Absolute Error (RAE) figures, which showed that Random Forest and Gradient Boost outperformed SVR and Bayesian Ridge Regression. When we switch to Dataset 2, the outcomes show different trends. Lower

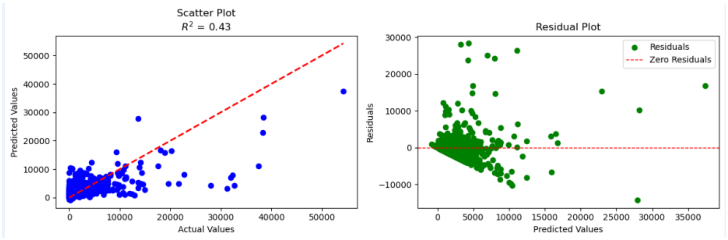

Figure 11: Synthetic data Gradient Boost regression scatter and residual plot

MAE values were shown by Decision Tree and Gradient Boost Regressor. Decision Tree demonstrated the lowest MAE, indicating superior accuracy. The negative R2 values for Random Forest and Gradient Boost indicate low predictive power, which may be caused by feature selection or dataset characteristics. Remarkably, Bayesian Ridge Regression showed significant R2 values, suggesting a superior fit for this dataset. The higher performance of Decision Tree in reducing relative mistakes is further highlighted by the RAE

values. Figure 11 and Figure 12 shows the scatter and residual plot of Gradient Boost for dataset 1 and Decision Tree for dataset 2 for regression tasks.

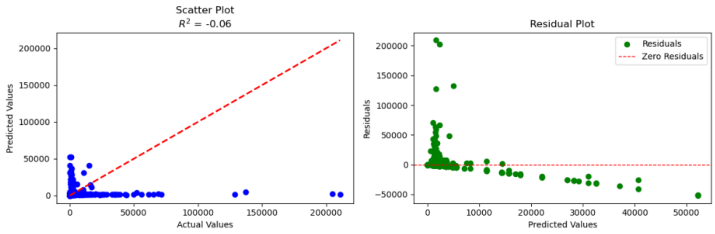

Figure 12: French MTPL Decision Tree regression scatter and residual plot

7 Conclusion and Future Work

To sum up, a thorough investigation into machine learning models to improve the precision of forecasting the frequency and value of auto insurance claims has produced insightful findings with useful applications. In response to the study topic, different classifier performance patterns were found in the experiments conducted on three different datasets. Remarkably, Random Forest and Logistic Regression showed resilience in multiclass classification, exhibiting flexibility in the face of unequal classes, but Adaboost encountered difficulties, especially in situations involving sampling. The outcomes of the binary classification confirmed Random Forest's consistently excellent performance across datasets.

Moving on to the regression studies, the results revealed subtle differences in model performance. In Dataset 1, SVR and Decision Tree Regression showed limits, but Gradient Boost and Random Forest predicted claim amounts with similar accuracy. Dataset 2 demonstrated Gradient Boost's improved predictive ability and highlighted its potential for accurate claim amount prediction. The discussion explored the intricacies of every model's functionality, taking into account measures such as MAE, R2, and RAE.

Implications for the insurance industry are significant, as the identified effective models, particularly Random Forest and Gradient Boost, can potentially revolutionize risk assessment and claim processing. The proposed future work emphasizes a meaningful exploration of tailored model extensions, delving into the specifics of class imbalances and dataset characteristics to further refine the predictive accuracy of machine learning models in the insurance domain. However, recognizing limitations in feature selection and dataset sizes, future endeavors could explore advanced techniques, such as ensemble methods and refined feature engineering, to enhance predictive capabilities. This research paves the way for practical implementations and underscores the continuous evolution of machine learning applications in optimizing insurance processes.

References

Alamir, E., Urgessa, T., Hunegnaw, A. and Gopikrishna, T. (2021). Motor insurance claim status prediction using machine learning techniques, *International Journal of Advanced Computer Science and Applications* **12**(3).

Arunkumar, N. and Yellampalli, S. S. (2017). Disruptive technology for auto insurance entrepreneurs: New paradigm using telematics and machine learning, *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, IEEE, pp. 195–200.

Baecke, P. and Bocca, L. (2017). The value of vehicle telematics data in insurance risk selection processes, *Decision Support Systems* **98**: 69–79.

Bian, Y., Yang, C., Zhao, J. L. and Liang, L. (2018). Good drivers pay less: A study of usage-based vehicle insurance models, **107**: 20–34.

Biau, G. (2012). Analysis of a random forests model.

Cunha, L. and Bravo, J. M. (2022). Automobile usage-based-insurance: : Improving risk management using telematics data, *2022 17th Iberian Conference on Information Systems and Technologies (CISTI)*, IEEE, pp. 1–6.

Dewi, K. C., Murfi, H. and Abdullah, S. (2019). Analysis accuracy of random forest model for big data – a case study of claim severity prediction in car insurance, *2019 5th International Conference on Science in Information Technology (ICSITech)*, IEEE, pp. 60–65.

Fauzan, M. A. and Murfi, H. (2018). The accuracy of XGBoost for insurance claim prediction, *International Journal of Advances in Soft Computing and its Applications* **10**(2): 159–171. Publisher Copyright: © 2018, International Center for Scientific Research and Studies.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H. and Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications, *Expert Systems with Applications* **73**: 220–239.

He, B., Zhang, D., Liu, S., Liu, H., Han, D. and Ni, L. M. (2018). Profiling driver behavior for personalized insurance pricing and maximal profit, *2018 IEEE International Conference on Big Data (Big Data)*, pp. 1387–1396.

Huang, Y. and Meng, S. (2019). Automobile insurance classification ratemaking based on telematics driving data, *Decision Support Systems* **127**: 113156.

Huang, Y. and Meng, S. (2020). A bayesian nonparametric model and its application in insurance loss prediction, **93**: 84–94.

Jeong, H. (2022). Dimension reduction techniques for summarized telematics data.

Liu, Z., Hao, M. and Tian, F. (2022). Ratemaking model of usage based insurance based on driving behaviors classification, **6**(2): 98–109.

McDonnell, K., Murphy, F., Sheehan, B., Masello, L. and Castignani, G. (2023). Deep learning in insurance: Accuracy and model interpretability using TabNet, **217**: 119543.

Noll, A., Salzmann, R. and Wuthrich, M. V. (2018). Case Study: French Motor Third-Party Liability Claims, *SSRN Electronic Journal* .

Peiris, H., Jeong, H. and Kim, J.-K. (2023). Integration of traditional and telematics data for efficient insurance claims prediction.

Poufinas, T., Gogas, P., Papadimitriou, T. and Zaganidis, E. (2023). Machine learning in forecasting motor insurance claims, **11**(9): 164.

So, B., Boucher, J.-P. and Valdez, E. A. (2021). Synthetic dataset generation of driver telematics.

Williams, A. R., Jin, Y., Duer, A., Alhani, T. and Ghassemi, M. (2022). Nightly automobile claims prediction from telematics-derived features: A multilevel approach, **10**(6): 118.

Yan, C., Wang, X., Liu, X., Liu, W. and Liu, J. (2020). Research on the UBI car insurance rate determination model based on the CNN-HVSVM algorithm, **8**: 160762–160773.

Zhang, Z. (2021). Data sets modeling and frequency prediction via machine learning and neural network, *2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT)*, IEEE, pp. 855–863.