Prediction of Insurance Complaints Settlements in Texas

Nabin G C
School of Engineering
Texas Tech University
Lubbock, USA
ngc@ttu.edu

Sahil Babu Bhetwal
School of Engineering
Texas Tech University
Lubbock, USA
sabhetwa@ttu.edu

Sriya Dhakal
School of Engineering
Texas Tech University
Lubbock, USA
srdhakal@ttu.edu

***Abstract*—The increasing number of insurance-related disputes in Texas creates an essential need for effective and transparent mechanisms to resolve the disputes. This project will leverage machine learning techniques to predict the outcomes of insurance complaint settlements, hence providing stakeholders with important insights into what drives resolutions. Using a comprehensive dataset comprising 259,694 rows and 17 columns, key attributes such as respondent roles, complainant roles, complaint reasons, and coverage types were analyzed. Advanced predictive models such as Logistic Regression, Random Forest, Support Vector Machines (SVM), LGBM Classifier, CatBoost Classifier, and Neural Networks were implemented and compared. Feature engineering techniques were done on the data, such as one-hot encoding for variables. After all said and done, the final model retained 11 critical features that achieved high accuracy in predicting settlement outcomes. This research underlines the potential role of machine learning for increasing transparency and fairness in insurance dispute resolution, to the benefit of insurers, policymakers, and consumers alike.**

I. Introduction

An insurance claim arises when a policyholder experiences a loss covered under their insurance policy and submits a claim to the insurer for compensation [1]. Insurance, in essence, is a contract in which one party agrees to provide financial compensation for specific losses in exchange for a premium paid by the other party. If a policyholder is dissatisfied with the coverage provided, they have the right to lodge a complaint against the insurance provider. The outcome of an insurance claim, known as the verdict, involves determining whether the claim is approved or denied based on the policy's terms and conditions. The growing number of insurance disputes in Texas highlights the need for effective and transparent mechanisms to resolve these conflicts. The Texas Department of Insurance handles complaints relating to individuals and organizations it licenses, such as insurance companies, agents, and adjusters, to ensure compliance with regulations for consumer protection in the insurance industry. [2]

The given project applies machine learning to forecast the outcomes of insurance complaint settlements and provides actionable insights for stakeholders into the key factors influencing resolutions. We began by conducting Exploratory Data Analysis (EDA) to identify underlying patterns, detect anomalies, generate hypotheses, and validate assumptions, with the goal of identifying a well-suited model. After the EDA the data were preprocessed. Categorical variables were encoded using one-hot encoding technique to ensure compatibility with the machine learning models. After preprocessing, the dataset contained 113,211 training records and 48,519 testing records, retaining the 11 most important features for model training. In the given work, advanced models are implemented on a dataset consisting of 161,000 records and 11 features: Among the used models: Logistic Regression, Random Forest, Support Vector Machines, LightGBM, CatBoost,Neural Networks, The CatBoost model achieved the highest accuracy at 81.43%, while the Support Vector Machines model had an accuracy of 79.31%. This research serves to underpin the potential of machine learning in promoting increased transparency, fairness, and efficiency in the resolution of insurance disputes.

II. Related Work

Insurance claim settlement and prediction processes have been extensively studied across various domains. A key area of research involves understanding the role of asymmetric information and its impact on settlement delays. Introduced a strategic model of pretrial bargaining in the context of personal injury claims, demonstrating how insurer beliefs, litigation costs, and liability assessments influence settlement timing. The study highlights that settlement delays are often influenced by factors such as high perceived case value, low bargaining costs, and reduced defendant liability beliefs. Despite these findings, the study emphasizes the need to incorporate two-sided informational asymmetries into dynamic pretrial models for better accuracy in explaining settlement delays. [3]

The use of machine learning (ML) for insurance prediction is another significant focus of research. Chukwuebuka and Zhen conducted a comparative analysis of multiple ML algorithms, including Logistic Regression, Decision Trees, K-Nearest Neighbors, Kernel Support Vector Machines, Naive Bayes, and Random Forests, for predicting building insurance claims. Through exploratory data analysis and model training on the Zindi Africa dataset, the study identified Kernel Support Vector Machines as the most accurate model, achieving 78% accuracy with a precision of 70.8%. Key variables such as building dimensions and types emerged as influential predictors. The findings underscore the importance of tailoring model

selection to the characteristics and quality of the data set for optimal predictive performance. [1]

Another perspective involves actuarial studies on the settlement of insurance claims. Traditional approaches often rely on Poisson-based models, as explored by Jiandong Ren, to estimate incurred losses and claim distributions. However, such models assume independence across settlement stages, which may not hold in real-world scenarios. To address this limitation, Markovian Arrival Processes (MAP) have been introduced as a more flexible framework capable of capturing dependencies across settlement stages and external environmental influences. This adaptability enables MAP to model complex claim behaviors, such as changing severities during settlement, offering a robust alternative to classical methods. Collectively, these studies contribute to a comprehensive understanding of insurance claims, from settlement dynamics to predictive modeling, paving the way for improved strategies in claim handling and management systems. [4]

III. Background

Machine learning (ML) is used to teach machines how to handle data efficiently, especially when extracted information from the data is difficult to interpret, and with the growing abundance of datasets, the demand for ML is on the rise. [5] Machine learning algorithm has its performance depends on the training success, dataset availability, data pre-processing, selection of attributes among others [1]. t involves the development of algorithms that can identify patterns in data, make predictions, and improve their performance over time based on experience. ML algorithms are typically divided into three main types: supervised learning, unsupervised learning, and reinforcement learning. The machine learning algorithms used in this project are:

Random forest: The Random Forest algorithm is an ensemble learning technique that can be used for both classification and regression problems. It was introduced by Leo Breiman in 2001 as a way to improve the performance of decision trees, which often suffer from overfitting and are not robust [6]. The basic idea of Random Forest is to combine many decision trees into a ”forest” and aggregate their predictions to improve the accuracy and reduce the variance. The algorithm predicts by taking an average of outputs over all the trees in the case of regression tasks, or by the majority vote in classification tasks. Random Forest has strong advantages to deal with large datasets featuring high dimensionality, missing value handling, and avoiding overfitting compared to a single decision tree. Random Forest is becoming one of the most popular, simple, versatile, and efficient algorithms for any machine learning problem, like feature selection, handling unstructured data, and prediction in complicated scenarios. [7]Robustness, ease of use, and good generalization on new, unseen data are among the reasons why it is one of the most used algorithms in machine learning.

CatBoost Another machine learning algorithm that is efficient in pre- dicting categorical feature is the CatBoost classifier. CatBoost is an implementation of gradient boosting, which makes use of binary decision trees as base predictors. [8] It uses an ordered boosting technique to avoid overfitting and givesvery high accuracy, especially in the case of small or imbalanced datasets. CatBoost is very scalable, performs parallel learning, and is generally great at classification, regression, and ranking tasks in machine learning.

Support Vector Machine: Support Vector Machines (SVM) are a supervised machine learning algorithm used for classification and regression tasks. SVM works by finding a hyperplane that best separates the data points of different classes in a high-dimensional space. Support Vector Machine a discriminative classifier is expressed by a separating line or hyperplane. [1]SVM is effective because it can handle both linearly separable and non-linearly separable data. For non-linear data, SVM employs the so-called kernel trick to transform the input data into higher dimensions for enabling linear separation. These kernels, which could be polynomial and RBF kernels, provide the versatility of SVM against complex data. [9]

Logistic Regression (LR) It is one of the most used algorithms for binary classification where the target variable can only be of two categories generally 0 and 1. Logistic regression is therefore the probability of a binary result depending on one or more predictor variables. Unlike simple regression of projecting continuous dependent variables, LR utilizes logistic (sigmoid) function to ensure that the value of predicted variable will range between 0 to 1, read as probabilities. The basic idea of LR is to estimate the model parameters (coefficients) that give the best fit of the observed to the predicted values through a process called maximum likelihood estimation (MLE) [10]. LR is preferred for its ease of use, speed and excess simplifications in areas such as healthcare, finance and marketing among other fields. However, this approach assumes that there is a linear trend where each predictor is directly proportional with log-odds of the outcome [11].

LightGBM It is a gradient boosting decision tree based model which is stably performing with the features of marking time and obtaining good results in high speed training, especially when dealing with large datasets. LightGBM introduces two key innovations: Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB). [12] Compared to the standard algorithm, GOSS is more efficient because it selectively considers only those data instances that have relatively large gradients while ignoring those data instances that have low gradients which do not significantly contribute to the calculation of information gains. In one way this makes the dataset smaller and likely to also increase the computational speed. To decrease the numbers to be processed and preserve sparsity of the feature vector, EFB merges “co-grouped” features where features are placed into several groups so that features in the same groups are mutually exclusive—features that are unlikely to have any value other than zero, usually not simultaneously. This methodology helps LightGBM to train much faster than regular implementations of the GBDT while preserving predictive efficacy consistently.

FeedForward Neural network Artificial Neural Networks, ANNs, are brief computational systems with architectural designs and structures that mimic biological neural networks. It composes steeped layers of nodes (neurons) or, in other words, it has visions that work in stacks in order to process data. Input, Weighted sum, Activation and Output are steps of a neuron; every neuron is a connection point that takes inputs, perform the weighted sum using an activation function, and then forward the result to the other layers. [13]An example of ANNs is the Feed Forward Neural Networks (FFNNs) which accept inputs and run through one or more hidden layers as well as an output layer while no connections are backward. Every neuron in each layer is connected with neurons in the subsequent layer and the computations are performed by passing inputs through an activation function weighed by the connection. FFNNs can be of the "single-layered' structure having only an output layer, or "multi-layered" possessing one or more layers of hidden neurons that can help the network learn features exhibiting order higher than two.

IV. MODEL DESIGN AND EXPERIMENTAL ANALYSIS

A. Model Design

The study evaluates six machine learning models—Logistic Regression, Linear SVM, Random Forest, LightGBM (LGBMC), CatBoost Classifier, and a Neural Network—on their ability to predict claim approval or denial during insurance conflict settlements using the dataset provided by the Texas Department of Insurance (TDI). All models were trained and tested on the same dataset, split in a 70:30 ratio, with k-fold cross-validation applied to ensure robustness. The data preprocessing steps included handling missing values, encoding categorical variables, and balancing the dataset, as detailed in Section IV-B.

B. Dataset

The dataset we used was collected from Texas Open Data Portal which consists of collected information by the Texas Department of Insurance (TDI) from 2012 to 2024. This dataset includes a row for each person and organization named in a complaint. Each row is a combination of: (1) a complaint number, which may be shared with another person or organization named in the complaint, and (2) a Respondent ID, which is unique to the person or organization named on that row. The variables in the dataset with their description are shown in Table I.

C. Tools used

Python libraries, including Scikit-learn, Pandas, Seaborn, and Matplotlib, were used for data preprocessing, visualization, and model implementation. Neural Network models were executed on the Keras/TensorFlow framework. SHAP values were employed for feature importance analysis, offering explainability and transparency for model predictions.

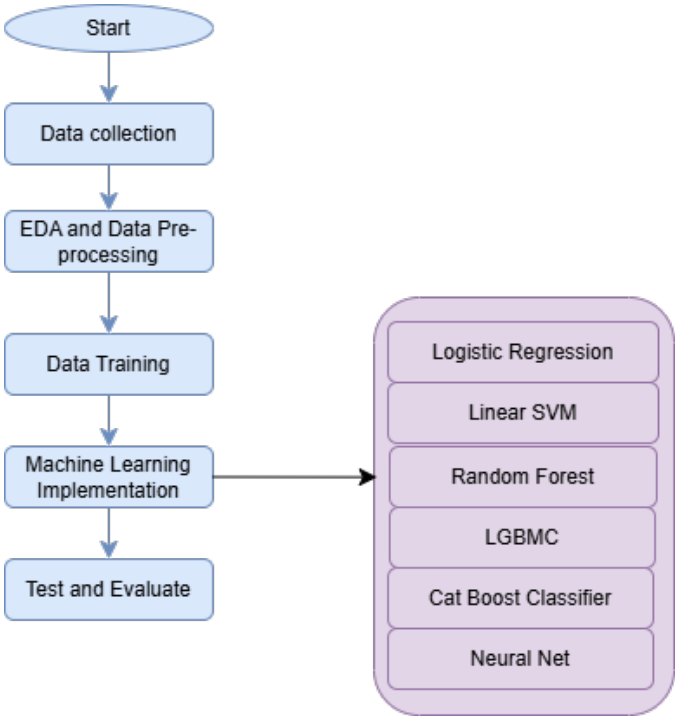

Fig. 1. Block Diagram of the Model

D. Data Preprocessing and Visualizations

After data collection the next step is exploratory data analysis followed by data preprocessing, integration and transformation. The most promising attributes of quality data include completeness, consistency, and timeliness.The performance of a mining algorithm depends on the quality of the data. But, the real-world data is incomplete and uncertain. Identification of the inconsistencies in the data is very difficult and even a very negligible amount of inconsistency in data degrades the performance of the mining algorithm at a very high rate.

Data visualization, on the other hand, aims to transmit the data clearly and effectively via graphical representation. Data visualization has found extensive use in many scenarios like reporting at work, task-progress tracking among others. One of the most popularly used advantages of the visualization technique is in discovering data relationships that may be hard to identify or observe by merely looking at the raw data.

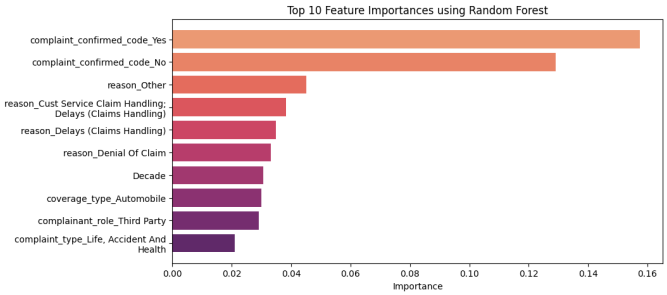

Fig. 2. Feature Importance Score

TABLE I
VARIABLES AND THEIR DESCRIPTION

No.	Variable	Description
1	Complaint number	The number assigned to a specific complaint.
2	Complaint filed against	The name of the person or organization the complaint was filed against.
3	Complaint filed by	Shows who filed the complaint (e.g., insured person, attorney, relative).
4	Reason complaint filed	Shows the reason the complaint was filed.
5	Confirmed complaint	Indicates if the complaint was confirmed (”Yes” means confirmed).
6	How resolved	A brief description of how the complaint was resolved.
7	Received date	The date TDI received the complaint.
8	Closed date	The date TDI closed the complaint.
9	Complaint type	Specifies if the complaint relates to ”Property and Casualty” or ”Life, Accident and Health.”
10	Coverage type	Indicates coverage, such as ”Automobile,” ”Homeowners,” or ”Miscellaneous.”
11	Coverage level	Specifies coverage level: ”Property or Casualty” or ”Life, Accident and Health.”
12	Others involved	Other parties involved, such as providers, attorneys, or hospitals.
13	Respondent ID	The ID assigned to the person or organization the complaint was filed against.
14	Respondent Role	The role of the person or organization the complaint was filed against.
15	Respondent type	Specifies if the complaint was against a person (individual) or organization.
16	Complainant type	Shows if the complaint came from a person (INDV) or organization (ORG).
17	Keywords	Shows more information about the complaint or category to help sort common issues.

We generated feature importance of the train data gini importance and Fig 2 shows the top 10 features. The correlation matrix is shown in Fig 3.

Since we have to design a system that will predict the probabilities of verdict for claim settlement against claim denial when an insurance conflict arises. The conflicts will be assessed based on the features and the target variable in this case being disposition.

1; if the verdict was settlement of claim
0; if the verdict was denial of claim

The preprocessed data were divided into a ratio of 7:3 training (70%) and testing (30%). The dataset was given a good description of the variables for each of the columns. Selection of the data or division of the dataset was done randomly and the models' performance was checked afterward. Since the data was abundant we removed all the rows where empty values were present. On both training and testing $complaint\_confirmed\_code$ and $involved\_party\_type$ included null values. Since the data was mostly categorical first the rare labels were combined to a category called 'Others' and then one hot encoded to fit the models.

The input parameters used for the experimental phase and purpose are shown in table II. We also included the k-fold with the best performance for each algorithm.

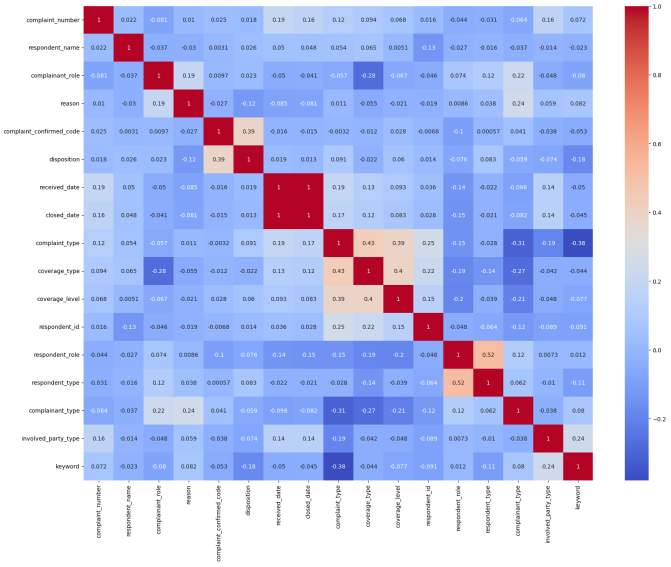

Fig. 3. Correlation Matrix

TABLE II
MODEL PARAMETERS AND SETTINGS

Model Name	**K-fold**	**Input Parameters**
Logistic Regression	10	Tolerance = 1×10^{-4}, Class weight = balanced, Solver = lbfgs, Random state = 0, Max iterations = 500
SVM	10	Kernel = linear, Random state = 0, Tolerance = 1×10^{-4}, Class weight = balanced, Max iterations = 1000
Random Forest	10	Number of trees = 100, Criterion = gini, Random state = 0
LGBMC	10	Boosting Type = gbdt, Random state = 0, Learning rate = 0.1, Number of estimators = 100, Number of leaves = 31
CatBoost Classifier	10	Max iterations = 200, Random state = 0, Learning rate = 0.2918

E. Evaluation Metrics

PERFORMANCE CRITERIA FOR ALGORITHM COMPARISON

The basis for the comparison of these algorithms anchors on some known performance criteria, some of which are outlined below with a short description for each of them. We choose these criteria because they are popular and easy to understand, in addition to the fact that they can be applied to all the algorithms for easier comparison.

Accuracy

This gives the estimate of the percentage of the actual rate values of all classes. A higher accuracy shows better performance (higher is better). Accuracy, simply put, is the ratio of appropriately predicted observations to the total obser-

vations. The formula for the calculation of accuracy is shown in Equation (1):

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

Precision

This determines the percentage rate of the true positive values for the relevant elements to the irrelevant ones. As with accuracy, higher precision percentages indicate that more relevant results were retrieved than irrelevant ones (higher is better). The formula for the estimation of precision is shown in Equation (2):

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Recall

Also referred to as Sensitivity Measure or True Positive Rate, recall is the fraction of a true positive rate of relevant values. A higher ratio means that more relevant elements were retrieved (higher is better). The method used for the estimation of recall is shown in Equation (3):

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

F1 Score

The F1 Score is a weighted mean of both precision and recall. The upper threshold value of the F1 Score is usually 1, which represents the best score, while the lower threshold value is 0, representing the worst score (higher is better). The formula for the estimation of the F1 Score is shown in Equation (4):

$$\text{F1 Score} = 2 \cdot \frac{\text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}} \tag{4}$$

Confusion Matrix

The result from the confusion matrix provides more detailed information about the performance of the algorithms by showing the number of truly predicted positive and negative outcomes. Below are the key terms:

a) True Positives (TP):: These refer to the correctly predicted values that are positive, meaning that the value of the class is actually "yes," and the model predicted the class as "yes." For instance, the model predicted that a house has a claim, and it actually does. The True Positive values estimated from the models are shown in Table IV.

b) True Negatives (TN):: These refer to the predicted negative values that were predicted correctly, meaning that the value of the class is actually "no," and the model predicted the class as "no." For example, the model predicted no claim when the house does not have any claim. The True Negative values estimated from the models are shown in Table IV.

c) False Positives (FP):: This is sometimes called a Type I error and occurs when the actual class is "no," but the model predicted it as "yes." For example, if the model predicts a claim when it is supposed to predict no claim. The False Positive values estimated from the models are shown in Table IV.

d) False Negatives (FN):: This is sometimes called a Type II error and occurs when the actual class is "yes," but the model predicted it as "no." For instance, if the model predicts no claim when it should have predicted a claim. The False Negative values estimated from the models are shown in Table IV.

Precision and Recall can all be calculated using the above parameters, as shown in Equations (1) to (4).

V. Results and Discussion

Our Project looked at the capability of different machine learning algorithm's ability to predict whether a claim is denied or approved during the verdicts of any insurance conflicts settlements according to the dataset provided by TDI. As we had enough data we split the data into 7:3 for all the models and used k-fold validation.

TABLE III
Performance Metrics for Models

Model Name	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.799	0.794	0.799	0.795
SVM	0.796	0.791	0.797	0.790
Random Forest	0.804	0.802	0.805	0.803
LGBMC	0.811	0.809	0.811	0.810
CatBoost Classifier	0.810	0.808	0.811	0.809
Neural Net	0.801	0.798	0.808	0.803

The results in table III illustrate the performance of six classification models—Logistic Regression, SVM, Random Forest, LightGBM (LGBMC), CatBoost Classifier, and a Neural Network—based on four metrics: Accuracy, Precision, Recall, and F1-score. The LightGBM model emerges as the top-performing model across all metrics, achieving an accuracy of 0.811, a precision of 0.809, a recall of 0.811, and an F1-score of 0.810. The CatBoost Classifier follows closely, with slightly lower but competitive values across the metrics, indicating its robust performance as well. Random Forest and Neural Networks perform reasonably well, with Random Forest showing balanced scores ($F1 = 0.803$) and Neural Networks demonstrating a strong recall value (0.808), suggesting their potential in scenarios requiring accurate positive class predictions. On the other hand, Logistic Regression and SVM, while achieving comparable accuracies (0.799 and 0.796, respectively), fall behind in other metrics, indicating their relatively lower suitability for the given task.

Among the models, the CatBoost Classifier's superior performance can be attributed to its ability to handle categorical features effectively and its boosting mechanism, which combines weak learners to achieve optimal predictions. The high scores across all metrics for this model suggest it is well-suited for both precision-oriented and recall-critical applications. LGBMC, another gradient boosting model, similarly benefits from its efficient training process and robust generalization capability, which is evident from its high and consistent metric values. Neural Networks and Random Forest offer strong alternatives, particularly in scenarios with high-dimensional

data or non-linear patterns, but their slightly lower scores suggest room for optimization. Overall, the results highlight CatBoost Classifier as the most reliable and effective model for the dataset, followed closely by LGBMC, with other models offering varied strengths depending on the application focus.

TABLE IV
CONFUSION MATRIX VALUES FOR MODELS

Model Name	TP	FP	FN	TN
Logistic Regression	9517	3735	6015	29350
SVM	8949	3287	6583	29798
Random Forest	10300	4255	5232	28830
LGBMC	10654	4307	4878	28778
CatBoost Classifier	10448	4127	5084	28958
Neural Net	10677	4338	4855	28747

The Neural Network model achieves the highest true positive rate $(10,677)$ and the lowest false negative rate $(4,855)$, suggesting it has the best recall performance among the models. However, it also has a relatively high false positive rate $(4,338)$, which could impact precision in some applications.

The CatBoost Classifier and LGBMC models also demonstrate strong performance, with high true positive rates $(10,448$ and $10,654$ respectively) and reasonably low false negative rates $(5,084$ and $4,878$ respectively). These models appear to strike a good balance between correctly identifying positive cases while maintaining a low number of false negatives. Logistic Regression and SVM have higher false negative rates $(6,015$ and $6,583$ respectively) compared to the other models, indicating they may struggle with correctly identifying positive cases to some degree.

Random Forest exhibits a moderate balance, with $10,300$ true positives and $5,232$ false negatives, making it a solid, albeit not the best, performer in this classification task. Overall, the confusion matrix highlights the trade-offs between the models' ability to correctly identify positive cases (recall) and their propensity for false positives (precision). The choice of the best model will depend on the specific requirements and priorities of the classification problem at hand.

The feature importance plot provides valuable insights into the key factors influencing the model's output. The most significant driver appears to be the absence of a complaint confirmation code, suggesting that this variable has a strong impact on the model's predictions. Additionally, the respondent's role as a self-funded ERISA plan and the type of insurance coverage (automobile) also emerge as important factors.

Several other features related to the reasons for claims handling, denial, and settlement offers, as well as the respondent's identity, further contribute to the model's performance. This comprehensive view of the influential variables allows for a deeper understanding of the underlying dynamics at play and can guide efforts to optimize the model and enhance its predictive capability. By identifying the most impactful factors, this analysis offers a solid foundation for refining the model and potentially improving the overall decision-making process.

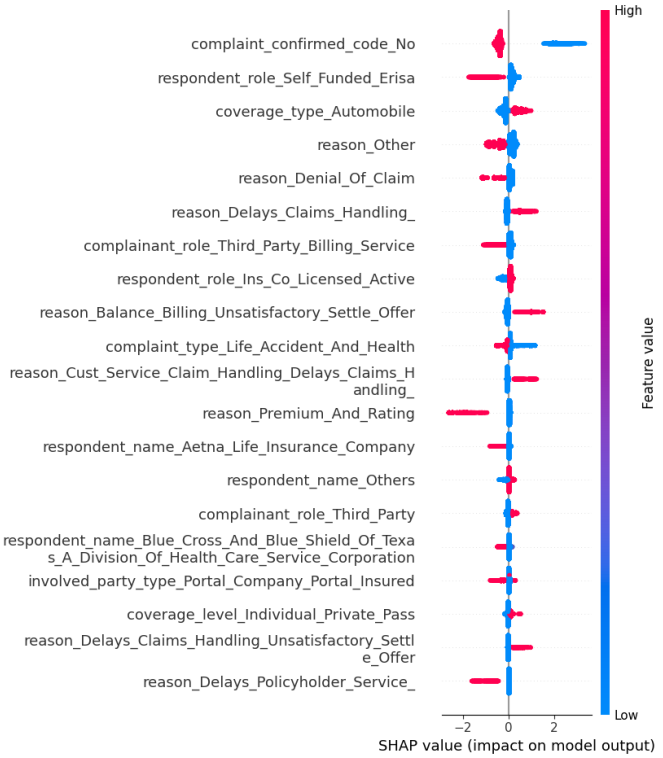

Fig. 4. SHAP Value

The limitations of the models we compared stem from the fact that their performance is inherently limited to the dataset on which they were trained, a common characteristic of most machine learning algorithms. However, a key strength of this analysis lies in its ability to assist practitioners in selecting the most appropriate model for a given scenario. By evaluating and comparing various models, the study provides valuable guidance for choosing the model that may perform most effectively under specific conditions.

One notable strength of the study is its application to a medium-sized dataset, which includes a substantial number of categorical variables. Despite the presence of widely distributed values and numerous anomalies within the data, the preprocessing techniques employed enabled the models to achieve acceptable levels of accuracy. This suggests that even in the presence of complex and imperfect datasets, it is possible to preprocess the data effectively and achieve reliable predictions. This finding underscores the potential of machine learning models to provide meaningful insights, even in challenging real-world scenarios such as insurance conflict resolution.

Additional considerations could include further exploration of model generalizability, particularly when applied to larger or more diverse datasets. It would also be useful to examine the impact of specific preprocessing techniques and feature engineering strategies on model performance to better understand the optimal conditions for each model.

VI. Conclusion and Future Work

The results highlight the importance of choosing models tailored to specific application needs. Future work could involve testing the models on larger, more diverse datasets and exploring advanced techniques to further enhance model performance and generalizability. This study lays the groundwork for implementing machine learning in insurance conflict resolution, showcasing the potential for data-driven decision-making. Additionally, expanding the analysis to include revenue-related claims data and predictions on settlements can provide a more comprehensive view of the financial impact of conflicts. Leveraging advanced Natural Language Processing (NLP) techniques, such as BERT or GPT, to analyze complaints can further refine insights, enabling a deeper understanding of customer grievances and enhancing resolution strategies. These approaches promise to drive more effective and efficient claim management in the insurance sector.

References

[1] C. J. Ejiyi, Z. Qin, A. A. Salako, M. N. Happy, G. U. Nneji, C. C. Ukwuoma, I. A. Chikwendu, and J. Gen, "Comparative analysis of building insurance prediction using some machine learning algorithms," *International Journal of Interactive Multimedia and Artificial Intelligence*, vol. 7, no. 3, p. 75, 2022. [Online]. Available: http://dx.doi.org/10.9781/ijimai.2022.02.005

[2] E. Kitzman, "Texas department of insurance," 2012.

[3] P. Fenn and N. Rickman, "Asymmetric information and the settlement of insurance claims," *The Journal of Risk and Insurance*, vol. 68, no. 4, p. 615, Dec. 2001. [Online]. Available: http://dx.doi.org/10.2307/2691541

[4] J. Ren, "Analysis of insurance claim settlement process with markovian arrival processes," *Risks*, vol. 4, no. 1, p. 6, Mar. 2016. [Online]. Available: http://dx.doi.org/10.3390/risks4010006

[5] B. Mahesh, "Machine learning algorithms - a review," *International Journal of Science and Research (IJSR)*, vol. 9, no. 1, p. 381–386, Jan. 2020. [Online]. Available: http://dx.doi.org/10.21275/ART20203995

[6] S. J. Rigatti, "Random forest," *Journal of Insurance Medicine*, vol. 47, no. 1, p. 31–39, Jan. 2017. [Online]. Available: http://dx.doi.org/10.17849/insm-47-01-31-39.1

[7] L. Breiman, *Machine Learning*, vol. 45, no. 1, p. 5–32, 2001. [Online]. Available: http://dx.doi.org/10.1023/A:1010933404324

[8] A. A. Ibrahim, R. L., M. M., R. O., and G. A., "Comparison of the catboost classifier with other machine learning methods," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, 2020. [Online]. Available: http://dx.doi.org/10.14569/IJACSA.2020.0111190

[9] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 1–27, Apr. 2011. [Online]. Available: http://dx.doi.org/10.1145/1961189.1961199

[10] D. W. Hosmer and S. Lemeshow, *Applied Logistic Regression*. Wiley, Sep. 2000. [Online]. Available: http://dx.doi.org/10.1002/0471722146

[11] D. R. Cox, "The regression analysis of binary sequences," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 20, no. 2, p. 215–232, Jul. 1958. [Online]. Available: http://dx.doi.org/10.1111/j.2517-6161.1958.tb00292.x

[12] M. Hajihosseinlou, A. Maghsoudi, and R. Ghezelbash, "A novel scheme for mapping of mvt-type pb–zn prospectivity: Lightgbm, a highly efficient gradient boosting decision tree machine learning algorithm," *Natural Resources Research*, vol. 32, no. 6, p. 2417–2438, Aug. 2023. [Online]. Available: http://dx.doi.org/10.1007/s11053-023-10249-6

[13] S. Murat H., "A brief review of feed-forward neural networks," *Communications Faculty Of Science University of Ankara*, vol. 50, no. 1, p. 11–17, 2006. [Online]. Available: http://dx.doi.org/10.1501/commua1-2_0000000026