*Auto Insurance Fraud Prediction*

Sun Han

Bellevue University

[DSC630-T301 Predictive Analytics (2255-1)](https://cyberactive.bellevue.edu/webapps/blackboard/execute/courseMain?course_id=_534952_1)

Frank Neugebauer

May 27, 2025

Milestone 5

**Introduction of Topic/Problem**

Insurance fraud costs U.S. consumers an estimated $300 billion annually, significantly impacting premiums. Auto insurance fraud is a growing concern, ranging from exaggerated claims to staged accidents. As fraudulent activities become more sophisticated, traditional detection methods often fail to identify and prevent these schemes. This research explores the potential of predictive analytics in detecting auto insurance fraud, aiming to enhance fraud prevention strategies and reduce financial losses. By leveraging this data-driven approach, insurers can improve fraud detection accuracy and protect honest policyholders from bearing the burden of inflated costs.

**Overview of Data Used**

The dataset carclaims.csv was sourced from <https://www.kaggle.com/datasets/khusheekapoor/vehicle-insurance-fraud-detection/data>. The dataset contains 15420 samples where 923 are fraudulent claims. This data has 33 features that provide details regarding the time of day, the accident location, the type of vehicle involved, policyholder demographics, claim history, and other relevant factors. Additionally, the dataset captures important behavioral and contextual indicators, such as the number of past claims, the age of the vehicle, the policyholder’s marital status, and the deductible amount. By analyzing these features, this study aims to identify patterns associated with fraudulent claims.

**Methods of Analysis**

Data Preprocessing

We began by importing the full carclaims.csv dataset, which contains 15,420 records and 33 features capturing claim details, policyholder demographics, and contextual factors. Our first step was to standardize missing‑value placeholders by converting any “0” or “none” strings in categorical columns to true NaN. Next, we mapped ordinal and binary fields to numeric codes, turning category labels (for example, policy type, vehicle category, and agent type) into ordered integers and mapping “yes”/“no” flags to 1/0. For remaining missing entries, we applied mode imputation for categorical features and filled numeric gaps with zero. After these imputations, we performed one‑hot encoding on any leftover nominal variables to ensure every column was numeric. A final quality check confirmed that no feature contained null values and that all data types were suitable for modeling. We then executed a stratified 80/20 train‑test split to preserve the roughly 6 percent fraud rate in both sets.

Exploratory Data Analysis

Visualizations play a crucial role in unraveling the nuances of our fraud detection data. For example, stacked bar charts reveal concentrated clusters by comparing fraud versus non-fraud counts across categorical variables such as policy type, accident area, vehicle category, and agent type, which highlights elevated fraud in "Sedan - Liability" claims and rural settings. Overlaid histograms, KDE plots, and box plots bring key numeric features like vehicle age, number of past claims, and deductibles into focus; they showcase differences in distribution between fraud and non-fraud cohorts while identifying concentrated segments. A correlation heatmap further elucidates which features tend to move together, even though the fraud label itself exhibits only weak linear correlations with individual predictors, suggesting that more intricate interactions may be at play. Additionally, ROC and precision-recall curves provide a dynamic view of model performance, illustrating the trade-offs between maximizing true positive rates and minimizing false alarms, and a bar chart of Random Forest feature importances succinctly ranks predictors to highlight the variables that carry the greatest predictive weight. This integrated suite of visual tools not only deepens understanding of the data’s structure but also guides strategic decisions in model tuning and risk management.

**A graph of a number of people

AI-generated content may be incorrect. A graph of a person and person

AI-generated content may be incorrect.**

**A screenshot of a graph

AI-generated content may be incorrect.A graph of a graph with text

AI-generated content may be incorrect.**

Analysis of the factors associated with fraudulent auto‐insurance claims reveals several clear patterns. First, when looking at policy types, most frauds are filed under “Utility – All Perils” and “Sedan – All Perils” coverages. In fact, when examining vehicle categories more broadly, sedans lead all other types in the number of fraudulent claims. Geographic location also plays a role: urban areas see far more fraud than rural regions. Demographically, male claimants are more frequently linked to fraud than female claimants, and single individuals report noticeably higher rates of fraudulent claims compared to those who are married or divorced. In terms of the base policy purchased, collision and liability policies are the settings in which fraud most commonly occurs. Finally, the involvement of external agents is correlated with a higher incidence of fraud, suggesting that internal agents may employ stricter verification procedures or oversight.

A graph of a number of data

AI-generated content may be incorrect.A graph of a number of supplements

AI-generated content may be incorrect.

Analysis of quantitative risk factors in fraudulent auto insurance claims highlights several notable trends. First, age plays a role: policyholders in their early thirties to early forties exhibit a slightly higher incidence of fraud compared to older groups. Secondly, vehicle value is inversely related to fraud frequency, as lower-priced cars, those valued roughly in the 1 to 2 range, see more suspicious claims than their higher-priced counterparts. Examining claim history, fraud peaks among those with no prior claims as well as those with three past claims, suggesting both first-time filers and repeat claimants at that level warrant closer scrutiny. Finally, the number of supplements submitted correlates with fraud, with the vast majority of fraudulent claims occurring among policyholders who submitted zero supplements, indicating that claims without follow-up documentation may be easier to manipulate.

A screenshot of a graph

AI-generated content may be incorrect.

Analysis of vehicle-related risk factors in fraudulent auto insurance claims reveals several clear patterns. First, older vehicles, those around six to eight years old, tend to be involved in fraudulent claims more often than newer models. Second, looking at claim history, fraud is notably more frequent among policyholders with zero, one, or three past claims, suggesting both first time filers and certain repeat claimants warrant extra scrutiny. Third, deductible levels show that most claims, including the fraudulent ones, cluster around a deductible of 400. Fourth, vehicle value again plays a role, as lower priced vehicles around a price level of two see higher numbers of fraudulent claims than higher priced cars. Finally, the number of supplements submitted correlates with fraud, as many fraudulent claims occur when no supplements are submitted, indicating that claims lacking follow up documentation may be easier to exploit.A group of blue boxes with white text

AI-generated content may be incorrect. Analysis of box-plot risk factors in fraudulent auto-insurance claims highlights a few key insights. First, the age of vehicles shows that fraud cases tend to involve slightly newer vehicles rather than older ones. Second, examining past number of claims reveals no significant differences between fraudulent and non-fraudulent claims. Third, both deductible amounts and vehicle price exhibit no clear disparities by fraud status. Finally, the number of supplements submitted does correlate with fraud, as fraudulent claims tend to have fewer supplements on average.

A screenshot of a graph

AI-generated content may be incorrect.

Analysis of feature correlations in fraudulent auto-insurance claims reveals that some variables, such as age of vehicle and age of policyholder, exhibit a moderately strong positive correlation of around 0.61. In contrast, the fraud indicator itself shows very weak correlations with all other variables. This suggests that no single feature has a strong linear relationship with fraud and that effective fraud prediction may require modeling more complex interactions or incorporating additional data sources.

Models Used

The models utilized in this study are logistic regression and random forest. These models were chosen due to their effectiveness in classification tasks and ability to handle structured data commonly found in fraud detection. Logistic regression is a baseline model that offers interpretability and insight into which factors contribute most to fraudulent claims. It helps understand the relationships between independent variables and the likelihood of fraud. Random Forest can improve prediction accuracy by combining multiple decision trees. It helps handle imbalanced datasets and capture complex patterns in fraud detection.

Evaluation Metrics

I plan to evaluate the results using precision, F1-Score, and Confusion Matrix. These metrics help handle the class imbalance in fraud detection. Precision measures the proportion of fraudulent claims classified as fraud, ensuring that the model minimizes false positives. F1-Score provides a balanced evaluation of the model’s performance, especially when false positives and negatives must be considered. The Confusion Matrix offers a detailed breakdown of the model’s predictions, showing the number of true positives, false positives, true negatives, and false negatives. This helps in understanding the model’s strengths and weaknesses in detecting fraud. To ensure robustness, I will also validate the model using cross-validation and a train-test split approach, ensuring that the results generalize well to unseen data. These evaluation techniques will help determine the most effective model for accurately identifying fraudulent insurance claims. In addition to precision, F1, and confusion matrices, we compute ROC AUC and Precision–Recall curves to assess trade‑offs between true positive rates and false alarms under class imbalance.

Model Building and Evaluation

Random Forest  
Our experimental pipeline centers on two Random Forest classifiers trained on the same feature set but differing in how they handle class imbalance. First, we fit a baseline Random Forest on the raw, imbalanced training data using default settings. Next, to address class imbalance, we applied SMOTE oversampling to synthetically balance the minority class before training a Random Forest model with class weighting (class\_weight='balanced').

For both models, we evaluated performance on the held-out test set by calculating precision, recall, F1 score, and ROC AUC. We also plotted ROC and precision-recall curves to visualize model performance. Additionally, confusion matrices were generated to inspect true and false positive rates side by side. Together, these metrics and visualizations provide a comprehensive comparison of how each approach trades off false positives against missed fraud.

Initial Model Performance (ROC/PR Curve)

A graph of a curve and a graph of a curve

AI-generated content may be incorrect.

Revised ROC/PR CurveA comparison of a graph

AI-generated content may be incorrect.

XGBoost Classifier

An XGBoost classifier was trained independently on the same preprocessed training data. Hyperparameters included use of the log-loss evaluation metric, and trees were grown to capture nonlinear patterns. No additional sampling or scaling steps were applied beyond the encoding and imputation performed earlier. A comparison of a graph

AI-generated content may be incorrect.

Stacking Ensemble with Logistic Regression, Random Forest, and XGBoost

In the final modeling step, all features were standardized using a StandardScaler. A stacking classifier was built with three base learners: the previously trained XGBoost model, logistic regression configured with an increased iteration limit, and a Random Forest composed of one hundred trees. A logistic regression meta-learner was trained on out-of-fold predictions from the base learners using five-fold cross-validation. This ensemble strategy combined linear and tree-based methods to capture complementary patterns in the data.

A close-up of a graph

AI-generated content may be incorrect.

**Results & Findings Explained**

Random Forest  
The baseline Random Forest model demonstrated limited performance in identifying fraudulent claims. Precision was moderately high (0.50), but the recall was extremely low, resulting in a very low F1-Score of 0.01. The confusion matrix revealed that the model correctly identified only one fraudulent case while missing 196, indicating that the model struggled significantly with recall.

After incorporating SMOTE to address class imbalance and using a class-weighted Random Forest, model performance improved dramatically. The revised model achieved perfect precision (1.00) and a recall of 0.94, yielding a robust F1-Score of 0.97. The confusion matrix showed that the model correctly identified 2,733 fraud cases while only misclassifying 167, indicating a significantly better balance between precision and recall. Additionally, cross-validation results demonstrated high consistency, with an average F1-Score of 0.90 and a standard deviation of 0.10.

The revised model's ROC AUC increased from approximately 0.86 to 0.99, reflecting its enhanced ability to differentiate between fraud and non-fraud claims. The Precision-Recall curve for the improved model was nearly perfect, emphasizing the model's reliable performance even under class imbalance.

XGBoost Classifier

The XGBoost classifier delivered outstanding performance on the test dataset. It achieved a precision of 0.99 and a recall of 0.96, resulting in an F1 score of 0.98. Examination of the confusion matrix revealed 25 false positives and 104 false negatives out of 5,799 observations. The ROC AUC value reached 1.00, demonstrating nearly perfect separation between fraudulent and legitimate claims. These metrics show that the gradient boosting framework effectively captures the complex nonlinear interactions within the features.

Stacking Ensemble

The stacking ensemble combined logistic regression, Random Forest and XGBoost to leverage their individual strengths. All features were standardized before training. Each base learner produced out-of-fold predictions which a logistic regression meta-learner then used under five-fold cross-validation. The ensemble obtained a precision of 0.99 and a recall of 0.97, yielding an F1 score of 0.98. Analysis of the confusion matrix indicated 31 false positives and 86 false negatives. Both the ROC curve and the precision-recall curve maintained near-ideal shapes across thresholds, confirming that blending linear methods with bagged and boosted trees produces the most consistent fraud detection results.

Findings

Our results indicate that while SMOTE-augmented, class-weighted Random Forest greatly improves recall, reliance on synthetic examples can introduce artifacts that may not reflect real fraud patterns. The XGBoost model achieved near-perfect discrimination without any synthetic sampling and therefore offers greater stability and faster inference for real-time scoring. Its ability to capture complex nonlinear interactions makes it a strong candidate for production deployment when consistent, low-latency predictions are required.

The stacking ensemble, which blends logistic regression, Random Forest and XGBoost, further benefits from combining native imbalance handling and diverse learning strategies. It is best suited for periodic batch scoring where additional computational expense is acceptable and where its aggregated strengths can recalibrate thresholds against emerging fraud schemes. In practice, real-time flagging should rely on the class-weighted XGBoost model, with human adjusters reviewing flagged claims to guard against any model bias. Batch evaluation using the stacking ensemble can validate and adjust decision cut-offs. A monitoring framework tracking probability distribution, false positive and negative rates, and shifts in feature distributions will ensure that any drift or bias is detected promptly. Continuous retraining on new claims and collaborative audits between data science and claims operations will maintain model effectiveness and fairness over time.

**Ethical Implications and Conclusion**Leveraging complex models such as XGBoost and stacked ensembles presents interpretability challenges because their decision structures can obscure the reasons why individual claims are flagged. This opacity makes it difficult for stakeholders to trace model predictions back to specific features and to understand the underlying reasoning. To mitigate these concerns, it is essential to report feature importances clearly and to document decision thresholds in accessible terms. Ultimately, this model must serve as a tool to assist human adjusters rather than a substitute for human judgment. Every high-risk classification should be reviewed by a real person to ensure that no automated decision stands alone and that claimants are treated fairly.

The use of synthetic oversampling techniques in developing the stacking ensemble introduces the risk that artificial patterns may not faithfully represent genuine fraud behaviors. Models trained on SMOTE-augmented data require ongoing validation against actual claim outcomes to detect any divergence between training assumptions and real-world cases. By restricting synthetic sampling to batch processes rather than applying it in real-time scoring, organizations can incorporate human verification before acting on flagged cases.

Continuous monitoring of performance metrics is critical to identify any drift in data distributions or emerging biases related to demographic or geographic factors. Periodic audits conducted by cross-functional teams, including data science, legal, and compliance, will help detect and correct disparate impacts. Transparent communication with policyholders about the role of automated fraud detection, together with accessible appeal procedures, will foster trust and accountability throughout the claims process.

Moving forward, integrating these models into a unified fraud detection pipeline will require careful calibration of probability thresholds based on investigation capacity and cost sensitivity. A live monitoring dashboard should track key performance indicators and signal when retraining is needed. Collaboration with claims adjusters will ensure that flagged cases receive human review and that feedback continues to improve model performance. Ongoing evaluation against fresh claim data will sustain detection accuracy as fraud tactics evolve.

This project demonstrates that selecting and configuring advanced predictive models can transform the way auto insurance fraud is detected. The class-weighted XGBoost model emerged as the premier choice for real-time scoring by achieving near-perfect discrimination without relying on synthetic examples. Its rapid inference and inherent handling of class imbalance make it well suited for flagging suspicious claims as they arrive. In parallel, the stacking ensemble combining logistic regression, Random Forest and XGBoost provides a powerful solution for periodic batch analysis when computational resources permit. Blending diverse learning algorithms in this way enhances stability and helps capture emerging fraud patterns. Together, these measures deliver a robust, data-driven fraud detection system that balances precision, efficiency and adaptability, with human oversight at its core.

**Reference**

* **Khusheekapoor. (n.d.).** *Vehicle Insurance Fraud Detection* [Dataset]. Retrieved from <https://www.kaggle.com/datasets/khusheekapoor/vehicle-insurance-fraud-detection>