ATENEO DE MANILA UNIVERSITY

Insurance Fraud Detection

Final Project Submission

In partial fulfillment of the requirements for the subject

CSCI 114 (Pattern Recognition)

Written by Group 3 of the CSCI 114 ST1 Class

Valdez, Juliana Ysabelle S.

Submitted to

Sir Paolo G. Dano

A.Y. 2024-2025

Table of Contents

About the Project

This project aims to utilize machine learning techniques to detect and prevent insurance fraud by analyzing claim data. By implementing various models, including **Decision Trees, K-Nearest Neighbors (KNN),** and **Autoencoders**, the project seeks to create a reliable system for identifying fraudulent claims. With a labeled dataset containing fraudulent and non-fraudulent claims, the system will improve fraud detection accuracy while minimizing false positives, supporting insurance providers in reducing financial losses and operational inefficiencies. This project highlights the transformative role of AI in enhancing security and trust in the insurance sector.

Background of the Problem

Insurance fraud is a pervasive issue, costing the industry billions of dollars annually. Traditional fraud detection methods rely heavily on **manual review and rule-based systems**, which can be inefficient and prone to errors.

Machine learning provides an opportunity to enhance fraud detection by identifying complex patterns in claims data that might elude manual processes. This project's objectives include building a system to classify claims as fraudulent or legitimate, leveraging advanced algorithms to optimize detection, and reducing false positives that burden legitimate customers.

Metrics and Goals

Metric: Fraud detection systems must strike a balance between identifying fraudulent claims (high recall) and avoiding misclassifications of legitimate claims as fraudulent (high precision). The following metrics will evaluate the model's performance:

Recall

Recall measures the proportion of actual fraudulent claims that the model correctly identifies. It is critical in fraud detection as failing to detect fraudulent claims can result in significant financial losses.

Goal: Achieve at least **80% recall** for fraud detection

- A high recall ensures that the system identifies most fraudulent claims. This minimizes the risk of missing fraud, which is crucial for reducing financial risks.

Precision

Precision calculates the proportion of claims flagged as fraudulent that are genuinely fraudulent. High precision minimizes unnecessary investigations of legitimate claims.

Goal: Achieve at least **70% precision** when detecting fraud.

- High precision ensures that the flagged claims are highly likely to be fraudulent, reducing wasted resources on investigating legitimate claims

F1-Score

The F1-score is the harmonic mean of precision and recall, balancing the trade-off between detecting as many fraudulent claims as possible and minimizing false alarms.

Goal: Achieve an **F1 score of at least 80%**.

- The F1-score ensures the model provides a balanced approach, prioritizing both effective fraud detection (recall) and reducing misclassifications (precision).

Scope and Limitations

The scope of this project is to develop machine learning models that analyze structured data (e.g., claim amounts, claimant history, and incident details) to detect insurance fraud.

The project focuses primarily on the development and evaluation of machine learning models for fraud detection, including preprocessing steps such as handling missing values, encoding categorical variables, and scaling numerical features. These steps were guided by existing methodologies from similar projects, ensuring consistency and efficiency in preparing the data for model training. However, this approach limits the scope of innovation in the preprocessing phase, as it heavily relies on prior work.

The dataset's limitations also affect the model's performance and generalizability. The quality and completeness of the dataset are central factors, with **missing or imbalanced data** potentially leading to biased models or overfitting.

Some categorical variables, such as policy_state, insured_occupation, and incident_type, have high cardinality, which may challenge the model’s ability to

generalize. The dataset also **lacks sufficient data** on certain types of fraud, limiting the models' ability to detect more complex fraud patterns.

Furthermore, this project does not explore alternative data sources or address the integration of these models into real-world claim processing systems. The use of unstructured data (e.g., text descriptions or scanned documents) is also excluded from the current scope, meaning the project does not leverage potentially valuable information that could improve fraud detection capabilities.

About the Dataset

Published in Kaggle, this dataset contains a comprehensive set of features related to insurance claims, focusing on various customer and incident attributes. It includes details such as customer demographics (e.g., age, sex, education, occupation), insurance policy characteristics (e.g., policy number, coverage limits, annual premium), and incident information (e.g., incident type, severity, collision details).

Additionally, the dataset features financial details about the claim, including total claim amount and specific claims for injury, property, and vehicle damage. A key attribute in the dataset is whether fraud was reported, which is essential for building predictive models for fraud detection. This dataset provides a rich source of data for analyzing factors contributing to insurance fraud and building machine learning models for automated fraud detection.

The new dataset undergoes significant transformation compared to the old one, with key steps including:

1. the removal of irrelevant or redundant columns like policy_number, policy_bind_date, and _c39;
2. the one-hot encoding of categorical variables such as policy_csl, insured_sex, insured_education_level, insured_occupation, incident_type, collision_type, incident_severity, property_damage, and police_report_available

These transformations convert object-type columns into multiple binary columns, making the data more suitable for machine learning models.

Additionally, the dataset addresses missing values by replacing categorical features with binary indicators (e.g., property_damage_NO and property_damage_YES), enhancing data completeness and making it more compatible with machine learning algorithms. More than this, as the dataset is labeled, it is suitable for supervised learning.

Selection of Models

Classical: Decision Tree

Decision trees are interpretable and can model nonlinear relationships, making them suitable for fraud detection where clear rules are beneficial for understanding patterns in the data. They are efficient and can handle mixed types of data (categorical and continuous), often present in insurance datasets (Quinlan, 1986). This model can handle overfitting through pruning techniques and is explainable to non-technical stakeholders.

Classical: K Nearest Neighbor

KNN is effective for detecting anomalies by measuring the similarity between instances in feature space. In insurance fraud detection, fraudulent cases often have distinct patterns compared to legitimate ones. KNN can classify such cases based on proximity to labeled examples. The primary strength of this model is the fact that it is simple and non-parametric, which means it does not make assumptions about the data distribution.

Modern: Autoencoder

Autoencoders are unsupervised learning models designed to detect anomalies by reconstructing input data. Fraudulent claims, being anomalies, often result in higher reconstruction errors. Insurance fraud detection can benefit from an autoencoder by identifying claims that deviate significantly from the norm. The strength of this model is in its ability to handle imbalanced datasets well.

In the context of fraud detection, autoencoders could be used as a form of **anomaly detection**—where fraud is treated as an anomaly, making it a hybrid approach. It may not directly use labeled data in the traditional sense like Decision Trees or KNN, but with labeled anomalies (fraud cases), you can apply it in a supervised manner by leveraging reconstruction errors for fraud detection.

On Model Development

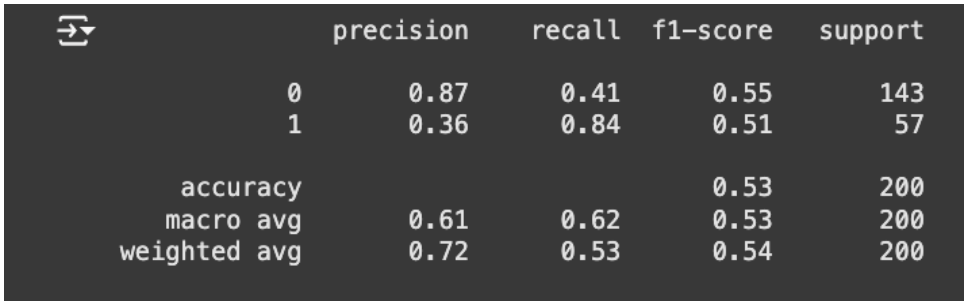
Trained Models

1. **K-Nearest Neighbors (KNN)**:
    - The KNN model was trained using the training data, where the model classifies data points based on their proximity to the nearest neighbors in the feature space.
    - **Cross-validation** was used to evaluate the performance of the KNN model. This involved splitting the data into multiple folds, training the model on a subset of the data, and testing it on the remaining data. This process was repeated several times to ensure a robust evaluation of the model's performance and prevent overfitting.
2. **Decision Tree**:
    - The Decision Tree model was trained on the dataset by learning decision rules based on the features. It recursively splits the data to create branches, optimizing for the best feature to split on at each level.
    - The model was trained using the training dataset, and the decision tree structure was learned to predict the target variable based on feature values.
3. **Neural Networks**:
    - The neural network model (Autoencoder) was trained on the data by minimizing the reconstruction error between the input and its predicted output. The encoder-decoder structure was used to learn a compact latent representation of the data and then reconstruct the data from this representation.
    - The model was trained using backpropagation, where the loss function (MSE) guided the weight updates, and the optimization was performed using stochastic gradient descent (SGD).

In summary, KNN utilized **cross-validation** for model selection and evaluation, while Decision Tree and Neural Networks were trained on the dataset using standard training procedures.

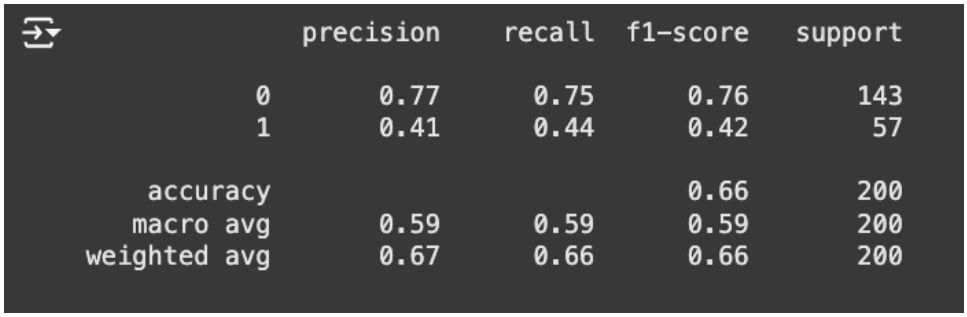
Results of Trained Models

Shown below are the classification reports of all the models.



	precision	recall	f1-score	support
0	0.87	0.41	0.55	143
1	0.36	0.84	0.51	57
accuracy			0.53	200
macro avg	0.61	0.62	0.53	200
weighted avg	0.72	0.53	0.54	200

Figure 01. Classification Report of Decision Trees



	precision	recall	f1-score	support
0	0.77	0.75	0.76	143
1	0.41	0.44	0.42	57
accuracy			0.66	200
macro avg	0.59	0.59	0.59	200
weighted avg	0.67	0.66	0.66	200

Figure 02. Classification Report of KNN

Discussion

The decision tree model has a high precision for normal claims (0.87), but its recall for normal claims is only 0.41. This shows that the decision tree model tends to classify most claims as normal, and misses many fraudulent claims (low recall). On the other hand, it has a high recall for fraudulent claims (0.84), indicating that it correctly identifies a large portion of fraud cases. However, its precision for fraud is only 0.36, suggesting that it falsely flags many normal claims as fraudulent. The overall accuracy of the decision tree is 53%, which is lower than that of the KNN model, and it seems to favor detecting fraud at the cost of misclassifying normal claims.

The KNN model shows a precision of 0.77 for normal claims (class 0), but only 0.41 for fraudulent claims (class 1). This indicates that the model is better at identifying normal claims, but it struggles with detecting fraud, as seen from its low precision for fraudulent claims. The recall for fraudulent claims is also low at 0.44, which means that a significant portion of fraudulent claims is not being detected by the model. The overall accuracy is 66%, with a macro

average recall of 59%, which indicates that the model is performing moderately well, but there is room for improvement, especially in detecting fraud.

Conclusion and Recommendations

Neither the KNN nor the Decision Tree model passed the desired fraud detection metrics. The KNN model exhibits a recall of 0.44 for fraudulent claims, which is significantly below the desired threshold of 80%, and its precision of 0.41 falls short of the 70% target. Similarly, the Decision Tree model shows a high recall for fraud (0.84), but its precision is only 0.36, leading to a high number of false positives. Additionally, both models suffer from overfitting: the KNN model struggles to generalize well, and the Decision Tree model demonstrates an imbalance in performance, with high recall for fraud but poor recall for normal claims, indicating it may be overfitting to the minority class (fraud). The lack of a proper balance between precision and recall in both models results in suboptimal performance. Given these issues, neither model met the minimum performance expectations for fraud detection. To enhance detection capabilities, it would be beneficial to explore further model optimization, adjust for overfitting, or consider more advanced algorithms such as ensemble methods or neural networks.

References

Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. Journal of Network and Computer Applications, 60, 19-31.

Chandar, R., & Alwan, A. (2019). High Cardinality Features in Machine Learning: A Practical Guide to Learning from Large-Scale Datasets. Springer.

C. S. Hodge, J. Austin, and P. C. Neumark (2007). Mining for Fraudulent Insurance Claims: A Survey of Methods, Challenges, and Applications. Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

He, H., & Wu, D. (2017). Imbalanced learning: foundations, algorithms, and applications. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 7(6), 252-274.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, J. V., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A review and a new framework. Expert Systems with Applications, 38(12), 14680-14690.