Insurance Complaint Analysis using Machine Learning Methods

Matthew Lucia
University of Vermont
Burlington, VT 05405
mlucia@uvm.edu

Conor McDevitt
University of Vermont
Burlington, VT 05405
cdmcdevi@uvm.edu

Ryan Courtney
University of Vermont
Burlington, VT 05405
racourtn@uvm.edu

Abstract

In this project, we analyze consumer insurance complaints using a machine learning pipeline aimed at classifying complaint disposition outcomes. We investigate various modeling techniques—including Logistic Regression, Random Forest, Neural Networks, and XGBoost—to evaluate their effectiveness in handling imbalanced and text-rich datasets. Preprocessing involved one-hot encoding categorical variables, normalizing numerical features, and leveraging pre-trained word embeddings for textual complaint descriptions. While Logistic Regression served as a basic baseline, both Neural Networks and XGBoost outperformed simpler models, achieving higher classification accuracy and more robust F1-scores across multiple classes. XGBoost, in particular, demonstrated strong performance in handling class imbalance and capturing non-linear feature interactions, while Neural Networks excelled in modeling text embeddings to extract nuanced semantic patterns from complaint narratives. Our results indicate that advanced models, when combined with thoughtful feature engineering and data balancing techniques such as SMOTE, can provide significant value in automating and improving the efficiency of complaint resolution workflows in the insurance domain.

1 Introduction and Problem Definition

Identifying common reasons for complaints can help insurance companies proactively address issues. Predicting complaint resolution time or recovery amount can assist in optimizing customer support, and analyzing complaint outcomes can provide insight into systemic inefficiencies.

Insurance companies receive many complaints regarding claim handling, policyholder service, underwriting issues, and more. Understanding the patterns in these complaints can help insurance providers improve their services, reduce disputes, and streamline resolution processes.

The primary objectives of our research include:

1. Developing predictive models to understand complaint characteristics
2. Creating embeddings to capture nuanced textual information
3. Improving complaint resolution processes through intelligent analysis

By building Neural Networks and Logistic Regression models with embeddings, we can extract meaningful features from complex complaint data, predict complaint dispositions and recovery amounts with higher accuracy, and provide insights into complaint patterns to assist insurance companies address issues.

CS 3540/CS 5540/CMPE 5990/CSYS 5990 Final Project (Spring, 2025)

2 Related Work and Bibliography

Since insurance claims are abundant and time-consuming to process, machine learning offers a promising solution. However, literature reflects no single agreed-upon approach. Text-heavy claims add complexity due to the subjectivity of language.

1. Models should be explainable. Insurance involves significant financial stakes, and stakeholders want to understand the reasoning behind predictions. Logistic regression offers interpretability, unlike some neural methods.
2. Gilpin et al. (2018) propose heuristic techniques including:
    - Local model-agnostic explanations (e.g., LIME, SHAP)
    - Visualization tools such as partial dependence plots and ICE plots

3 Model and Training Algorithm

1. **Inputs**
    - Categorical: Coverage, SubCoverage, Reason, SubReason, Status
    - Numerical: Complaint duration, Recovery amount
    - Text: Conclusion Statement
2. **Model Architecture**
    - Text Embeddings: Pre-trained (e.g., Word2Vec)
    - Neural Network: Embeddings $\rightarrow$ dense layers $\rightarrow$ softmax
    - Logistic Regression with embeddings and L1/L2 regularization
3. **Outputs**
    - Classification: Complaint reason, disposition
    - Regression: Recovery amount, resolution time

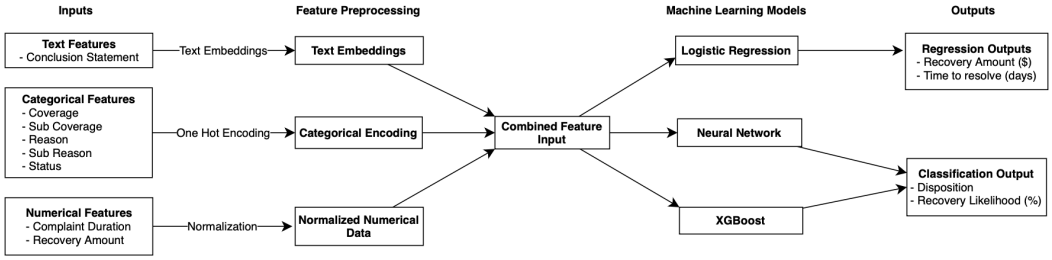

Figure 1: Pipeline

4 Dataset

4.1 Data Source

"Insurance Company Complaints" dataset from Kaggle. Public dataset of consumer complaints filed in Connecticut.

4.2 Dimensions

- 32,267 rows, 12 columns

4.3 Data Preprocessing

- Impute incomplete rows/columns
- Derived features:
  - Complaint duration: Days between 'Closed' and 'Opened'
  - Text embeddings
- Normalize numerical features
- Encode categorical variables (one-hot)

4.4 Embedding Strategy

Pre-trained embeddings (Word2Vec, GloVe) to represent text features semantically.

5 Experimental Evaluation

5.1 Evaluation Methodology

To assess the performance of our models, we used a combination of loss metrics, classification reports, and confusion matrices. These tools provide both high-level and granular insights into how well the models performed on the classification task.

The classification report includes standard performance metrics such as accuracy, precision, recall, and F1-score. Accuracy provides an overall measure of model correctness, while precision and recall offer a deeper understanding of the model's behavior in distinguishing between specific classes — particularly in identifying false positives and false negatives. The confusion matrix further supports this analysis by visually representing the distribution of predicted versus actual labels.

As a starting point, we implemented two models of varying complexity: logistic regression random forest. These models served as a baseline for comparison. Logistic regression was chosen because it is a simple, interpretable model that provides a useful reference point. While it lacks the complexity and capacity of more advanced models like neural networks and decision trees, its performance establishes a lower bound against which improvements from more sophisticated models can be measured.

This model performance can be compared to the random forest classification model, which is slightly more sophisticated and complex. Comparing the results from these two models tells us if our data benefits from a more complex model, and gives us insights into the dataset.

5.2 Results

1. Logistic Regression (baseline)
   - accuracy: 0.34

   This metric suggests the model is not performing very well. This is expected, as a logistic regression model is much too simple to capture the fine details and correlations in the data.
   - Low F-1 score for smaller categories, and High F-1 score for large categories.

   This shows that the model rarely makes correct predictions for classes with fewer samples, and the high recall indicates there are many false positives in the classifications. This is reflected in the low F-1 score for these classes.
2. Random Forest
   - accuracy: 0.60

   This metric shows a significant improvement over the Logistic Regression model, which indicates that the data benefits from a higher degree of model complexity.
   - Low F-1 score for smaller categories, and High F-1 score for large categories.

This trend continues from the Logistic Regression Model, suggesting that the models are likely not predicting the disposition correctly, but rather are classifying most data points into the most common category, giving the illusion of correct predictions.

This indicates problems that need to be addressed in training more complex models. One techniques considered is using SMOTE (Synthetic Minority Over-Sampling Technique) to helpe combat the class imbalance in the dataset.

5.3 Discussion

Our initial experimentation with baseline models of Linear Regression and Random Forest Classification show two main things:

- The data benefits from a more complex model to capture fine details and correlations between features.
- The data is severly imbalanced, with a few categories dominating the dataset, which is an issue that needs to be addressed when training subsequent models.

6 Conclusion

To be completed.

References

[1] Johnson, M., Albizri, A., & Harfouche, A. (2023). Responsible Artificial Intelligence in Healthcare: Predicting and Preventing Insurance Claim Denials. *Information Systems Frontiers*, 25, 2179–2195. https://doi.org/10.1007/s10796-021-10137-5

[2] Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An approach to evaluating interpretability of machine learning.