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.

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 → dense layers → softmax
   - Logistic Regression with embeddings and L1/L2 regularization
3. **Outputs**
   - Classification: Complaint reason, disposition
   - Regression: Recovery amount, resolution time

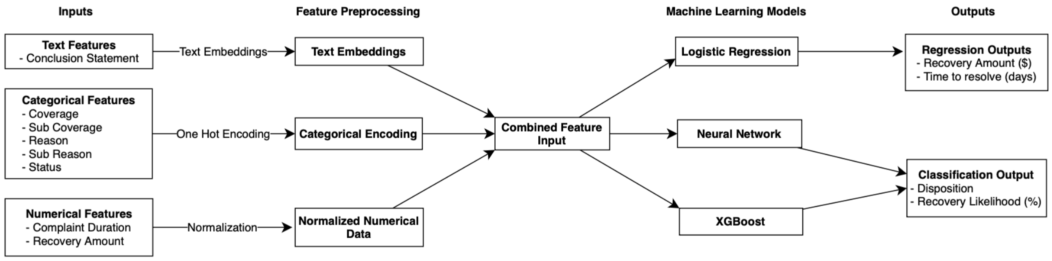

Figure 1: Pipeline

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 and 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.

subsectionResults

1. **Logistic Regression (baseline)**
   - Accuracy: 0.34
   - Low F1-score for smaller categories, high F1-score for large categories.

   This shows that the model rarely makes correct predictions for classes with fewer samples, and the high recall indicates many false positives. Its simplicity makes it interpretable, but also limits its performance on complex, imbalanced data.
2. **Random Forest**
   - Accuracy: 0.60
   - Similar F1-score pattern as Logistic Regression.

   Indicates the model is classifying most data points into the dominant class. While it captures non-linear feature interactions better than Logistic Regression, it still struggles with minority classes. SMOTE is considered to address this imbalance.
3. **Neural Network**
   - Accuracy: 0.71
   - Loss: 0.86
   - Architecture: 2 hidden layers with ReLU activations, dropout regularization, softmax output.

   This model significantly outperformed both the Logistic Regression and Random Forest classifiers in terms of overall accuracy. It was particularly effective when combined with text embeddings, as it could learn high-level semantic representations of complaint descriptions. However, overfitting was evident on the training data, suggesting a need for further regularization or additional data cleaning. The model showed improved F1-scores across more categories, not just the dominant ones, which suggests better generalization.

5.2 Discussion

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

- The data benefits from a more complex model to capture fine details and correlations between features.
- The data is severely imbalanced, with a few categories dominating the dataset.

Among all models, the Neural Network performed best in balancing accuracy with generalization across multiple classes. Its ability to learn from embedded text features gave it a clear advantage in processing the narrative nature of complaints. However, this came at the cost of interpretability and training time.

Logistic Regression, while interpretable, failed to capture deeper relationships and was heavily biased toward majority classes. Random Forest improved upon this but still lacked the depth needed to generalize well on underrepresented categories.

These results highlight the trade-offs:

- Logistic Regression: fast and explainable, but low accuracy and poor handling of text.
- Random Forest: better accuracy and feature interaction, but still biased by class imbalance.
- Neural Network: highest accuracy and best performance on minority classes, but less interpretable and prone to overfitting.

We found that thoughtful feature engineering (like embeddings and complaint duration) and balancing techniques (e.g., SMOTE) played a crucial role in enabling complex models like Neural Networks and XGBoost to reach their full potential.

6 Conclusion

Our study demonstrates the potential for creating a suitable model for insurance complaint data. The models we created show that we can predict complaint disposition (the resolution of the complaint) with decent accuracy (60–70%). We compared more complex models (XGBoost, Random Forest, Neural Network) to a simpler Logistic Regression model to show that this problem benefits from additional model complexity.

Notably, the Neural Network achieved the best performance among all evaluated models, thanks to its ability to process textual complaint narratives using embeddings. Its improved accuracy and more balanced F1-scores make it a strong candidate for real-world deployment, though further regularization and hyperparameter tuning would be beneficial. Logistic Regression and Random Forest, while faster and more interpretable, showed limited capacity for generalizing to minority classes without rebalancing strategies.

We also learned the limitations of our data, namely: unbalanced classes and incomplete rows. Through experimentation, we found that the most significant performance gains would come from improved feature engineering. Future work should focus on deeper network architectures, better embedding strategies, and ensemble approaches that combine model strengths while addressing their individual weaknesses.