Homework Group 17 - Describe the overall structure of your final report

Introduction

This project addresses the task of automating decisions on compensation in insurance claims related to workplace injuries. Using historical data from the New York WCB, our objective is to predict the type of compensation a claim will be awarded, thereby optimizing the decision-making process and improving consistency in outcomes.

This project aims to build a robust predictive model to classify injury claims in a multiclass classification scenario, explore and optimize model performance through preprocessing techniques and hyperparameter tuning and provide interpretative analysis to identify the most influential variables in the final decision.

Prior studies show that machine learning techniques, such as random forests and XGBoost, have proven effective in automating decision-making processes in insurance. Our work builds on these approaches, exploring multiple variables related to demographics and injury types to create an interpretable and reliable decision-making model.

Data Exploration and Preprocessing

Data Description

The provided data includes information on injury characteristics, such as type, location, claimant demographics, hearing records, and prior agreements. Key observations reveal that claimants predominantly fall between 30 and 60 years of age, with weekly wages displaying a wide range, including extreme values that could influence the model's outcomes. Additionally, certain injury locations, such as the lower back, appear more frequently, indicating a higher risk associated with specific occupations.

Preprocessing Steps

We started by handling Missing Values. After checking the missing values, we deleted the column, *OIICS Nature of Injury Description,* because it had no content. We removed and stored *WCB Decision, Agreement Reached*, later on we will create a model that predicts these variables. The records from the missing values were removed from the critical columns, such as *Claim Injury Type*, to maintain data integrity for analysis. For numerical variables with missing values, we used the median for imputation, while in categorical variables, we imputed with the most frequent category to avoid statistical biases.

In the outlier treatment, we identified and retained certain extreme values to ensure the model accurately reflects real-world data, particularly where extremes indicated strange values for age, birth years or average wage.

For variable conversion, we transformed categorical columns to category types, adjusted date columns to datetime types, and treated variables like "Claim Identifier" as strings to prevent misinterpretation.

Finally, we applied MinMax Scaling to the numerical data to ensure balanced influence of all features on the predictive models. For the categorical features, we applied an Ordinal Encoder to prepare the data for model input. For the target variable, "Claim Injury Type," we used a Label Encoder to convert it into numerical form suitable for the modeling process.

Multiclass Classification

For the multiclass classification model, it was necessary to perform a Feature Selection, combining correlation analysis and variable importance from an initial Logistic Regression model. Thus, features retained with high predictive value include Age at Injury, Industry Code, and Average Weekly Wage. Model Assessment Strategy and Metrics.

Diogo Ruivo, 20240584
José Tiago, 20240582
Matilde Miguel, 20240549
Nuno Sousa, 20222125
Rafael Lopes, 20240588

We tested several algorithms, including XGBoost, Logistic Regression, Decision Tree, Random Forest and SVM, chosen for their robustness and suitability for multi-class classification. To ensure reliable performance evaluation, we apply K-Fold cross-validation as our evaluation strategy. Finally, for performance metrics, we focus on accuracy, F1 score, and precision, placing special emphasis on macro F1 score.

Performance Comparison

The Decision Tree model exhibited the best initial performance in terms of accuracy and generalization, followed by XGBoost after applying some feature selection. Results are summarized in a comparison table for easy reference.

In our optimization efforts, we focused on fine-tuning the hyperparameters of the Random Forest and SVM models using Grid Search and Randomized Search techniques. For Random Forest, we adjusted parameters such as the maximum depth and the number of trees, while for SVM, we optimized the C parameter. These adjustments were aimed at improving the model's ability to handle complex patterns and enhance overall performance.

Following the optimization process, we observed a significant improvement in model accuracy, with an increase of approximately 5-10%. Notably, XGBoost outperformed the other models, proving to be the most reliable and effective algorithm for the multiclass classification task we were tackling. This result highlights the importance of hyperparameter optimization in achieving better model performance and more accurate predictions.

Open-Ended Section

We will work to improve the accuracy of predictions for the type of claim by exploring related variables and optimizing models. Our goal will be to identify key factors that impact model performance and understand limitations, especially for more complex injury classes.

First, we will test models to predict related variables such as "Agreement Reached" and "WCB Decision", stored, to determine whether their inclusion can improve predictive results, which will require analyzing correlations and potential causal effects. The Random Forest and XGBoost models will be selected and optimized through Grid Search, focusing on computational efficiency. Tuning hyperparameters will improve accuracy and consistency, especially with XGBoost. Other models will be tested to ensure a comprehensive performance comparison. Finally, we will perform K-fold cross-validation to check the robustness of the models, along with error analysis to identify areas with predictive difficulties.

We suspect that XGBoost will outperform other models in terms of accuracy and consistency. Variable analysis should reveal that some secondary variables contribute, albeit modestly, to improving forecast accuracy. Error analysis will highlight that certain classes of claim types will be more difficult to predict, pointing to areas where future techniques and feature engineering can improve the accuracy of these classes. Additional results demonstrated that Average Weekly Wage and Industry Code were strongly predictive of certain classes, such as Temporary or Permanent. The Agreement Reached variable also showed an indirect relationship with benefit types, indicating possible biases in decision-making.

By combining these approaches, we will aim to improve the prediction process and refine future strategies to address challenges related to prediction accuracy and class imbalance.

Conclusion

The primary objective of building an accurate predictive model for Claim Injury Type classification was successfully met. Key predictors included demographic and injury-specific factors like Age at Injury and Industry Code, confirming the significance of these features in outcome predictions. The results aligned with previous research, emphasizing the importance of contextual variables. Future improvements could involve exploring advanced models like neural networks.