A Critique of ‘Risk Prediction in Life Insurance Industry Using Supervised Learning Algorithms’

Nimalan Subramanian

March 22, 2024

Introduction

In the field of life insurance, risk assessment is a crucial component in classifying applicants. The underwriting process is used to make decisions on applications and price policies. Due to the rising number of applicants for life insurance, many insurance companies seek a faster, automated process to classify applicants and make decisions. Through the use of supervised learning algorithms, automating such a process has been proven to be both possible and effective. However, this process also has various concerns, particularly ethical considerations. This stems from a sense of privacy in the data and whether the process itself is fair. As such, the breakdown for how these algorithms are made and utilized must be studied to determine whether they are morally sound to implement.

Summary of Method

Data

Data was collected from various online databases and made into a final data set of 128 variables and 59,381 observations (applications) that describe the characteristics of each life insurance applicant. The data set is made of nominal, continuous, and discrete variables that are anonymized. The data acquired from Prudential Life Insurance particularly required a significant amount of cleaning to treat a large amount of missing values in order to make the data consistent.Variables with more than 30% missing data was dropped, resulting in only 4 attributes (Employment_Info_1, Employment_Info_4, Employment_Info_6, and Medical_History_1) serving as features retained for further analysis.

The data was then tested for the Missing Completely At Random (MCAR) mechanism (the distribution of missing values do not show any relationship between the observed data and the missing data), using Little’s test (null hypothesis was that the missing data was MCAR). The test resulted in a significance value of 0.000, implying that the null hypothesis was rejected, further revealing that the missing data was not entirely random. As such, the data can be either Missing at Random (MAR -> the missingness is dependent on other observed variables, but independent of unobserved features) or Missing Not At Random (MNAR -> the missing pattern is reliant on unobserved variables).

To determine which missing mechanism data is at hand, patterns in data set were examined through a tile plot for the missing values, with variables having the most mossing values at the top of the y-axis and the least missing values at the bottom. This visualization siggested that there is a random distribution of missing observations, leading the data set being assumed to be MAR. In this case, multiple imputation was used to appropriately replace the missing values, using available data to predict missing values in three steps: Imputation, Analysis, and Pooling. Imputation was done through the Multivariate Imputation via Chained Equations (MICE) package in R, following the removal of categorical variables and only the use of numerical variables. Following the evaluation of parameter estimates and standard erros through analysis, ther results are then integrated into a final result (pooling).

Learning Algorithms

To make modeling more efficient, the data set underwent dimensionality reduction through feature selection and feature extraction. Feature selection selects the prominent variables in the set, whereas feature extraction transforms higher dimensional data into fewer dimensions to build models. This was done to train machine learning algorithms faster and increase model accuracy by reducing the chance of overfitting. Four machine learning algorithms based on Correlation-based Feature Selection (CFS) and Principle Components Analysis Feature Extraction (PCA) were created to accomplish this. CFS grades subsets of the variables based on a hypothesis that is a useful subset of features that contain highly correlated features with the class that are not also correlated to each other. Using CFS removed noisy data and improved the performance of the learning algorithms. PCA extracts features that contain the most information to create new features (principal components), which are then used as new variables to be utilized in the prediction model.

Using Waikato Environment for Knowledge Analysis (WEKA), the CFS was implemented through a BestFirst search method on a subset evaluator, leading to 33 variables being selected out of a total 117 features, excluding the response variable. The PCA was implemented through a Ranker search method on a PrincipalComponents evaluator, which provided a rank for all 117 variables, then chose the optimal variables for prediction based on standard deviation to combine and create new features for the target variable prediction. Variables that had a standard deviation that was half of the first principal component (2.442) were kept, resulting in 20 variables (standard deviation of 1.221 or more).

The reduced data set was then utilized to build prediction models using the following machine learning algorithms: Multiple Linear Regression, REPTree, Random Tree, and Multilayer Perceptron. Multiple linear regression showed the relationship between the response variable and at least two predictor variables by fitting a linear equation to the observed data points. A Reduced Error Pruning Tree (REPTree) used regression tree logic to create numerous trees in different iterations, serving as a fast learner that develops decision trees based on the information gained and the variance reduction. The algorithm hooses the best of several created trees through the lowest MSE when pruning the trees. Random trees are then made based on random selection of data and variables to conduct backfitting, estimating class probabilites based on a hold-out set. Finally, the Multilayer Perceptron is used for backpropogation to calculate the error of each neutron adter a subset of the data is processed, distributing the errors back through layers of an artifical neural network that can be altered during training.

Results

The above models were utilized on the CFS and PCA. For the CFS, the REPTree model showed the highest performance with the lowest mean absolute error (MAE) of 1.5285 and lowest root mean square error (RMSE) of 2.027. For the PCS, the multiple linear regression model yielded the best performance, having lower MAE and RMSE values of 1.6396 and 2.0659, respectively. In both instances, the random tree models has the highest error values. When comparing between feature selection and feature extraction, CFS models overall had lower errors compared to PCA. When considering algorithm performance, multiple linear regression, REPTree, and random tree had better results with CFS, while artificial neural network methods had a better performance with PCA.

Normative Concern

While these algorithms are efficient for companies to evaluate the risk levels of insurance patients, there are a number of ethical concerns with this process. The primary issue at hand stems from the ambiguity of data privacy and consent. While the process details that a step includes ensuring anonymity of its applicants. However, this information is still accessible prior to data cleaning and manipulation. Even in the case that such information is not provided, the primary variables that remain, employment history and medical history, can still be traced back to the applicant. Due to this concern, the purpose of consent is considered. While applicants may comply to the standards of consent provided by the company, the consideration of informed consent v tacit consent arises. Are applicants truly provided with all information that is digestible to them and are they providing their consent knowing everything they are complying to? In the case of tacit consent, potential misuse of data can occur, and the true benefits of providing consent must be explored.

Apart from data privacy and consent, an argument of fairness arises in the process to determine the risk prediction for each applicant. Justice as equality versus merit serves as a major point of consideration. In this field, merit would be used as the main factor, exemplified by the use of employment history and medical history as the primary variables to consider. On the other side of this argument, are these variables a fair measure of providing that value to the applicants? Without such considerations, the reliance on these variables could potentially result in a form of algorithmic bias that favors certain types of applicants over others. Overall, a consideration of the true degree of consent and the extent to which this process is fair must be done.