Research Paper selected for Synopsis

Risk prediction in life insurance industry using supervised learning algorithms

Boodhun, N., Jayabalan, M. Risk prediction in life insurance industry using supervised learning algorithms. *Complex Intell. Syst.* **4**, 145–154 (2018). https://doi.org/10.1007/s40747-018-0072-1

1. Problem Statement

Risk prediction in life insurance industry using supervised learning algorithms to classify the applicants and price policies accordingly

2. Paper Contribution

- Proposed solutions to enhance risk assessment among life insurance firms using predictive analytics
- Demonstrated executive dashboard using Microsoft Power BI for dynamic visualizations to gain better insights about the data
- Demonstrated the use of dimensionality reduction to select most viable attributes
- Consolidated the comparison of multiple Machine Learning algorithms like, Multiple Linear Regression, Artificial Neural Network, REPTree and Random Tree with Correlation-based feature selection (CFS) and Principal components analysis (PCA) feature extraction techniques. Also concluded and proposed, the importance of machine learning algorithms to efficiently predict the risk level of insurance applicants.

3. Dataset

Dataset used in the research paper is gathered from Prudential Life Insurance. The data set has 59,381 instances and 128 attributes.

Attributes	Type	Description
Product_Info_1-7	Categorical	7 normalized attributes concerning the product applied for
Ins_Age	Numeric	Normalized age of an applicant
Ht	Numeric	Normalized height of an applicant
Wt	Numeric	Normalized weight of an applicant
BMI	Numeric	Normalized Body Mass Index of an applicant
Employment_Info_1-6	Numeric	6 normalized attributes concerning employment history of an applicant
InsuredInfo_1-6	Numeric	6 normalized attributes offering information about an applicant
Insurance_History_1-9	Numeric	9 normalized attributes relating to the insurance history of an applicant
Family_Hist_1-5	Numeric	5 normalized attributes related to an applicant's family history
Medical_History_1-41	Numeric	41 normalized variables providing information on an applicant's medical history
Medical_Keyword_1-48	Numeric	48 dummy variables relating to the presence or absence of a medical keyword associated with the application
Response	Categorical	Target variable, which is an ordinal measure of risk level, having 8 levels

Figure 1: Dataset description

4. Data Pre-processing

Data pre-processing, which involves data cleaning, leads to Identification of outliers. All the outliers are removed from the target dataset along with suitable imputation method for missing values.

4.1. Missing Data

For missing data, three mechanisms are highlighted in the paper. After using the below mentioned mechanisms, the attributes that are showing more than 30% missing data are identified. These attributes are considered to be dropped.

4.1.1. Missing Completely At Random (MCAR), where distribution of missing values do not show any relationship between observed data and missing data, i.e. missing values are like a random sample of all the cases in the feature

4.1.2. Missing At Random (MAR), where missing data may be dependent of other observed variables, but independent of any unobserved features. i.e. missing values do not depend on the missing data, yet can be predicted using the observed data

4.1.3. Missing Not At Random (MNAR), which implies that the missing pattern relies on the unobserved variables, i.e. observed part of the data cannot explain the missing values

Dataset is tested for MCAR using the Little's test. Null hypothesis (missing data are MCAR) is rejected with a significance value of 0.000.

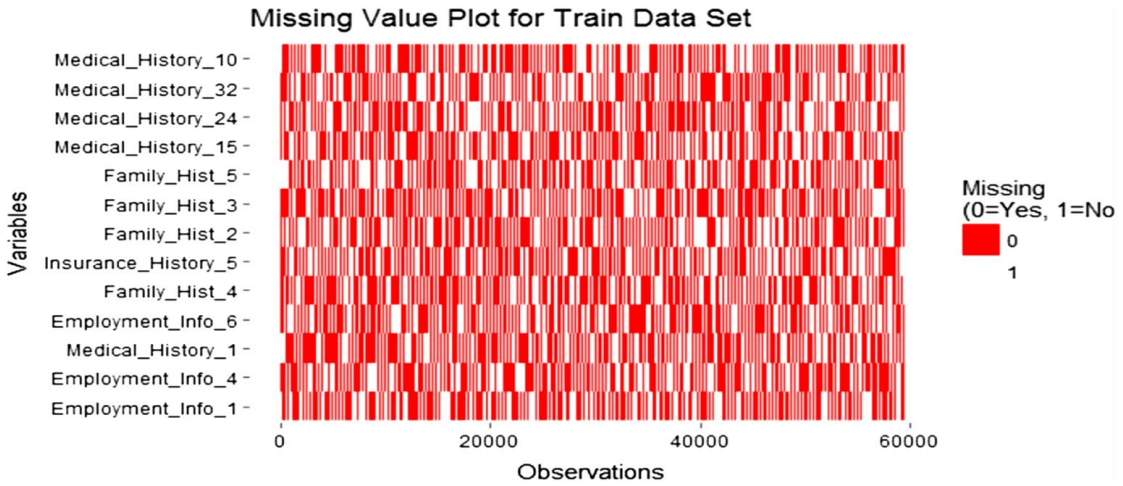

Figure 2: Missing value plot for train data

(Variable having most missing values on the top of the y-axis and least missing values on the bottom)

The visualization of the missing data structure suggests a random distribution of the missing value observations. The pattern of missing data and non-missing data is scattered throughout the observations. Therefore, the data set in this study is assumed to be MAR.

MICE (Multivariate Imputation via Chained Equations) has been utilized to do the multiple imputations. Categorical variables were removed and only numeric attributes are used to do the imputation.

4.2. Dimensionality Reduction

Dimensionality reduction involves reducing the number of variables to be used for efficient modeling which can be broadly divided into:

4.2.1. Feature selection (selecting the prominent variables) – Filter methods, wrapper methods, and embedded methods. Attributes can be selected based on Pearson's correlation, Chi-square or information gain ratio (IGR)

4.2.2. Feature extraction (transform the high dimensional data into fewer dimensions) – Correlation based feature selection (CFS) method and Principal component analysis (PCA) based feature extraction method.

PCA was implemented using a Ranker search method on a Principal Components, attribute evaluator. Cut-off threshold of 0.5 has been used to decide on the number of principal components to retain from the data set i.e., only those attributes which standard deviation value that is half of that of the first principal component (2.442) are retained resulting in 20 attributes only. Similarly, 33 attributes are selected by the CFS method.

5. Exploratory Data Analysis (EDA) / Visualization

In the paper, EDA is used by the researcher for the following:

- To understand different distributions that the features exhibit
- For bivariate analysis – Relationship between different features and Response attribute i.e. Risk level
- To understand the extent to which the independent variables are capable of impacting the response variable significantly

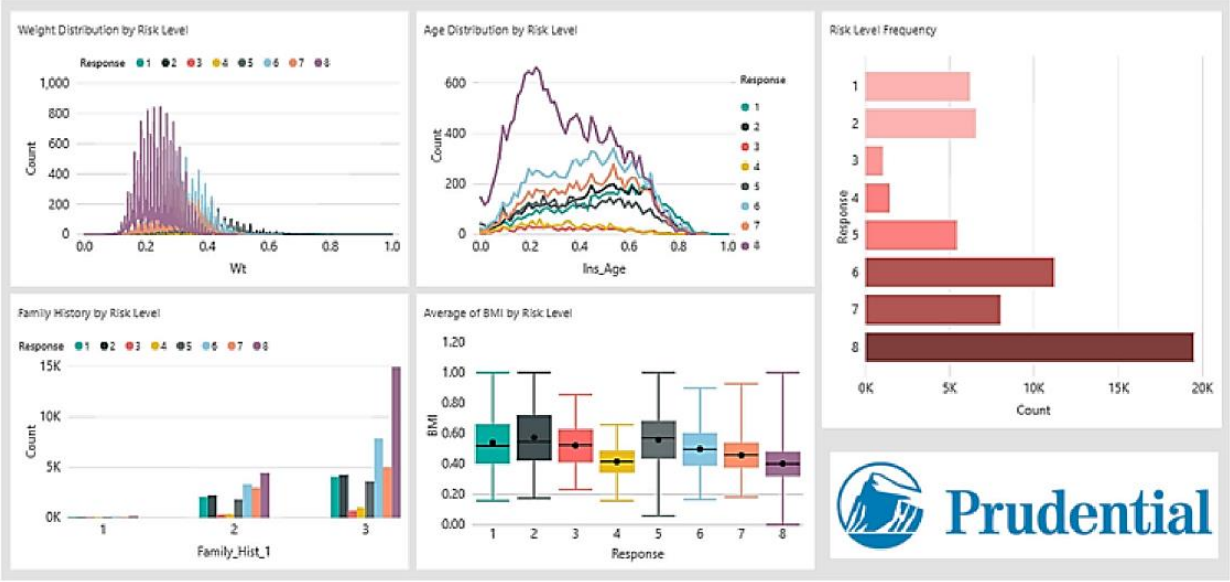

Figure 3: Life Insurance Executive Dashboard

6. Machine Learning Activity

Supervised machine learning is utilized to build the predictive models. Four techniques are explored in the paper – Multiple Linear Regression, REPTree, Random Tree, and Multilayer Perceptron

6.1. Multiple linear regression shows the relationship between the response variable and at least two predictor variables by fitting a linear equation to the observed data points. The variables significance in the regression equation are determined by statistical calculations and are mostly based on the collinearity and partial correlation statistics of the explanatory features

6.2. REPTree (Reduced Error Pruning Tree) classifier is a type of decision tree classification technique which can build both classification and regression trees, depending on the type of the response variable. It makes use of regression tree logic to create numerous trees in different iterations based on information gain and variance reduction. After creating several trees, the algorithm chooses the best tree using the lowest mean-square-error measure when pruning the trees

6.3. Random tree classifier builds a decision tree based on random selection of data as well as by randomly choosing attributes in the data set. Unlike REPTree classifier, this algorithm performs no pruning of the tree. In the research this algorithm is used together with CFS and concluded that the classifier works efficiently with large data sets.

6.4. Artificial neural network is an algorithm, which works like the neural network system in the human brain. It comprises of neurons organized in input, hidden and output layers. It utilizes back propagation to classify instances

Model validation has been performed using a k-folds (tenfold) cross-validation.

7. Result Analysis with Metrics

Algorithms	Error measures			
	CFS		PCA	
	MAE	RMSE	MAE	RMSE
Multiple linear regression	1.5872	2.0309	1.6396	2.0659
Artificial neural network	1.7859	2.369	1.7261	2.3369
REPTree	1.5285	2.027	1.6973	2.1607
Random Tree	1.7892	2.7475	2.0305	2.9142

Figure 4: Comparison of algorithms between CFS and PCA

For the CFS, the model developed using REPTree classifier shows the highest performance with the lowest mean absolute error (MAE) value of 1.5285 and lowest root mean square error (RMSE) value of 2.027 as compared to the other models.

However, for the PCA, the model developed with multiple linear regression shows the best performance with the lowest MAE and RMSE values as 1.6396 and 2.0659, respectively. Moreover, random tree classifier shows the highest error values for both feature selection techniques.

Comparing between the feature selection and feature extraction techniques, CFS shows that most of the models achieved lower errors compared to PCA. Multiple linear regression, REPTree,

and random tree classifiers show better performance when used with CFS, while artificial neural network shows a better performance with PCA.

8. Conclusion

Data analytics is now the trend that is gaining significance among companies worldwide. In the life insurance domain, predictive modeling using learning algorithms can provide the notable difference in the way which business is done as compared to the traditional methods.

From this research paper, we learnt the KDD approach to machine learning along with multiple metrics to understand the performance of ML algorithms which can be used in the field of Insurance.

Overall, the research paper contributes to the understanding of the factors that influence the purchase of life insurance policies and provides insights that can help insurance companies, policymakers, and researchers to develop more effective strategies to predict effective price of the insurance policies.

Thank You