Risk Prediction Assessment In Life Insurance Company Through Dimensionality

Reduction Method Sandeep Kumar Dwivedi, Ashish Mishra, Rajeev Kumar Gupta

Abstract: Risk assessment is one of the major components in life insurance organization through which customers are grouped. These type of life insurance organization has to perform different operations so that they can settle on different choices bases on applications and to keep proper management. But nowadays there is major expansion in data collection due to large number of customers and advances in investigation process. This is the reason these analysis process has been automated for faster process. Through this automation process many updation can be done although it also helps to include the different new plans by predictive analysis approach. Although real world dataset consist of large numbers of features that are used for examination, that's why dimensionality reduction has been applied to pick the selective attributes or features by which the power of the model can be increased. The dimensionality reduction can be done by strategies like Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), Correlation-Based Feature Selection (CFS), etc. Various machine learning classification methods like Artificial Neural Network, Multiple Linear Regression, Random Tree and the proposed Random Forest are applied on the dataset to predict the risk level of candidates. This work has shown that Backward Elimination Calculation has shown the most prominent result with least root mean square error (RMSE) OF 0.384 using the random forest strategy. This paper has also shown the training accuracy and testing accuracy on the basis of Random forest model.

Index Terms: Big data, PCA, RMSE, classification, backward elimination, random forest, feature selection

1 INTRODUCTION

The big data technology alter the manner in which insurance agencies collect, process, break down, and manage data more efficiently [1][2]. This technology has done remarkable improvement in different areas of insurance industry, for example, risk assessment, client investigation, item advancement, promoting examination, claims examination, guaranteeing investigation, extortion reinsurance and recognition [3]. Telematics is an another model where big data examination is in effect endlessly actualized and is changing the manner in which auto wellbeing net suppliers esteem the premiums of individual drivers [4]. Single life insurance associations still rely upon the customary actuarial equations to anticipate death rates and premiums of life techniques. These companies has doing judicious assessment to improve their business adequacy, anyway there is so far a nonattendance of expansive research on how prescient examination can advance the life coverage space. Specialists have focused on information mining frameworks to recognize fakes among insurance firms, which is an essential issued to the organizations confronting extraordinary misfortunes [5]. Manulife protection organization in Canada was the first to offer assurance to HIV suffering competitors through separating survival rates. Examination helps in the embracing method to give the benefit premiums to the right peril to keep up a vital good ways from troublesome assurance. Farsighted examination has been used by Property and Casualty (P&C) security net suppliers for over 20 years, on a very basic level for scoring inadequacy states on the probability of recovery. Risk profiles of individual candidates are altogether examined by guarantors, particularly in the life coverage business. The activity of the guarantor is to ensure that the dangers are assessed, and premiums as precisely as conceivable to continue the smooth running of the business.

- *Sandeep Kumar Dwivedi is currently pursuing masters degree program in Computer science and engineering in SISTec Bhopal, India, . E-mail: sandeep200391@gmail.com*
- *Dr. Ashish Mishra is currently Assistant professor in Computer science and engineering in JIIT Noida, India. E-mail: ashishmishra81@mail.com*
- *Dr. Rajiv Gupta is currently Associate professor in Computer Science and engineering at SISTec Bhopal,India . E-mail: rajivgupta@sistec.ac.in*

Risk classification is a typical term utilized among insurance agencies, which alludes gathering clients as indicated by their assessed degree of dangers, decided from their recorded information. Failure in distinguishing these risk factors can also create an issue referenced before known as adverse selection. There are two potential approaches to manage adverse selection issues:

1. Builds up an authoritative measure which makes an impulse measure available
2. Computation of risk on an individual basis. The obligatory buy of life insurance policy in the market by the administration makes Life Insurance Company too enormous to come up short, and consequently they can handle adverse selection. Although these type of method cannot be implemented in developing country, people might protest against government on implementing this policy. Another way is the estimation of requested risk for every policyholder from lots of characteristic and behavioral information. Proposed framework comprehends these basic risk factors and test whether they can add to recognizing essential features. This framework improves execution time of any machine learning model by distinguishing measurably huge highlights without trading off the extensive exactness of the model. Expansive research has not been done around there. The explanation behind this investigation is to apply predictive modeling to describe the risk level subject to the open past data in the insurance agency and recommend the most fitting model to access risk and offer responses for refine embracing structures.

1.1 Motivation

As the given writing demonstrates that low guaranteeing limits are a noticeable operational issue among insurance agencies overviewed anyplace. Another danger to the life insurance organizations is that they can confront unfriendly choice. Unfavorable determination alludes to a circumstance where the back up plans don't have all data on the candidate, and they wind up giving life coverage arrangements to clients with a high-risk profile. Protection firms with skilled guaranteeing groups weight on making the least potential misfortunes. As it were, the safety net providers endeavor to keep away from antagonistic choice as it can impactsly affect the life insurance business. Unfriendly choice can be maintained a strategic distance from by accurately grouping the hazard levels of

individual applications through prescient investigation, which is the objective of this examination. The exploration approach includes the accumulation of information from online databases. The theories about potential connections between factors would be explored utilizing defined intelligent advances. The exploration worldview manages a positivist methodology, as it is for the most part a prescient report including the utilization of machine learning to help the research objective. Risk evaluation is a urgent component in the life business to order the candidates. Organizations perform endorsing procedure to settle on choices on applications and to value approaches in like manner. With the expansion in the measure of information and advances in information investigation, the guaranteeing procedure can be mechanized for quicker handling of uses. This exploration goes for giving answers for improve chance evaluation among disaster protection firms utilizing prescient investigation.

2 LITERATURE SURVEY

Throughout the years, life insurance organizations have been endeavoring to sell their items proficiently, and it is realized that before an application is acknowledged by the life coverage organization, a progression of assignments must be embraced during the endorsing procedure. According to Wuppermann [4] endorsing includes gathering broad data about the candidate, which can be a long procedure. The candidates for the most part experience a few therapeutic tests and need to present all the pertinent records to the insurance agent. At that point, the financier surveys the hazard profile of the client and assesses if the application should be acknowledged. Therefore, premiums are determined. Prince stated that whether non-exposure of unasked for hereditary data comprises misrepresentation and investigates changing kinds of protection addresses that could possibly be translated as looking for hereditary data. Life insurance candidates by and large have no obligation to uncover unasked for data, including hereditary data, on an application. Be that as it may, given the complexities of genetic data, people might be presented to misrepresentation and rescission of their extra security arrangement in spite of legitimate endeavors to honestly and totally answer all application questions. Mamun et. al. has shown an impression of the seriousness and probability of the issues and prospects of the insurance business from the perspective of the insurance agencies themselves. The investigation uncovered that low capability of the specialists to be the most squeezing human asset the executives issue while the absence of specialized representatives remained as the most significant operational issue. Clients' absence of comprehension of protection terms and approaches and undesirable challenge turned out to be the most extreme advertising and moral issues individually. Timothy et.al has analyzes uneven data in the extra security market utilizing information that connection life coverage property with death records for an agent test of buyers. This examination finds no convincing proof for unfriendly determination in an expansive age accomplice. Hedengren and Stratmann [5] has observed unfriendly choice hypothesis predicts individuals with a high danger of death are bound to possess life coverage. Utilizing a one of a kind informational index combining regulatory and study records, they test this hypothesis and locate the inverse: individuals with high demise hazard are more averse to claim disaster protection. They propose beneficial determination and value separation swamp unfriendly choice in individual life coverage markets, where insurance agencies can't cost segregate just as in individual markets. Carson et.al. [6] Build up a model of protection valuing under heterogeneous slip by rates with unbalanced data about pass probability inside the setting of a discretionary two-section duty as a screening gadget for future policyholder conduct. At that point done test for shopper self-choice utilizing nitty gritty, approach level information on extra security predating (a typical practice that takes after a two-section levy). They can recognize, through a control capacity approach, the data about slip by risk a customer uncovers when they predate. Octaviani and Devi [7] have venture plans to find portfolio bunches by utilizing k-means grouping calculation and concentrate the principles of each group by creating order model utilizing Decision Tree calculation. The consequence of the model demonstrates that the bunches give various qualities and conduct. Supplement with KPI measurements, the organization can screen the exhibition of every group. So that the organization may utilize the investigations to enhance the technique of development and benefit. Noorhannah Boodhun and Manoj Jayabalan [8] has explored giving answers for improve chance appraisal among disaster protection firms utilizing prescient investigation. This present reality dataset with more than hundred qualities (anonymized) has been utilized to direct the investigation. The dimensionality decrease has been performed to pick noticeable traits that can improve the expectation intensity of the models. The information measurement has been decreased by highlight choice procedures and highlight extraction to be specific, Correlation-Based Feature Selection (CFS) and Principal Components Analysis (PCA).

3 PROPOSED WORK

In the proposed system the predictors are the features extracted from the dataset. They are treated as a null hypothesis. This will imply that there is no relationship between dependent variable and one or more combination of independent feature variable(s). Rejecting or disproving the null hypothesis gives a ground to believe that there is a significant relationship between dependent and independent variable(s). Whereas accepting the null hypothesis only provides the insignificant relationship between them. Hence, it is better to remove predictors if they are incapable to contribute a significant increase in accuracy of any machine learning model. If not removed, then their presence will only increase the execution time of machine learning models and will make model worst. So, it becomes important to understand this relationship by running test of the Null hypothesis on some method by which we can have a guarantee to reduce type I error with a goal of making the power of test close to one. One such method is known as the multiple linear regressions [9]. In the proposed work we will use a dataset that is given by Life Insurance Company and that is further used for analysis purpose so that risk assessment can be done. During implementation we have done dimensionality reduction so that feature can be reduced. Dimensionality reduction is needed because our dataset contains almost 128 features which is highly dimension. The high dimension dataset are considered to be very complex so dimensionality reduction is needed. There are two main approaches in dimensional reduction which are feature selection and feature extraction [10]. Our research work is based on feature selection.

3.1 Feature Selection

Feature Selection - Feature selection methods work more like filters that eliminate some attributes. There are on a very basic level two sorts of highlight choice procedures: filter type and wrapper type. Filter methodologies work by choosing just those characteristics that position among the top in gathering certain expressed criteria. Wrapper methodologies work by iteratively choosing, by means of an input circle, just those characteristics that improve the presentation of a calculation. In this we have taken wrapper based technique and that is named as backward elimination method [13]. In the fig 1 we have shown the flow chart of proposed work.

Description of Flow Chart:

Step 1: Data Preprocessing:

- Real-world data are large and inconsistent, deficient and ailing in certain behaviour and patterns. Thus preprocessing is needed very much.
- Data goes through series of steps during preprocessing like data cleaning, information coordination, information change, data reduction, data discretization and dealing with absolute illustrative factors in regression models.

Step 2: Set p-values and check the assumption of multiple linear regression for all predictors.

- A Regression test like linearity, normality, errors independence, multicollinearity lacking should also be checked before proceeding.

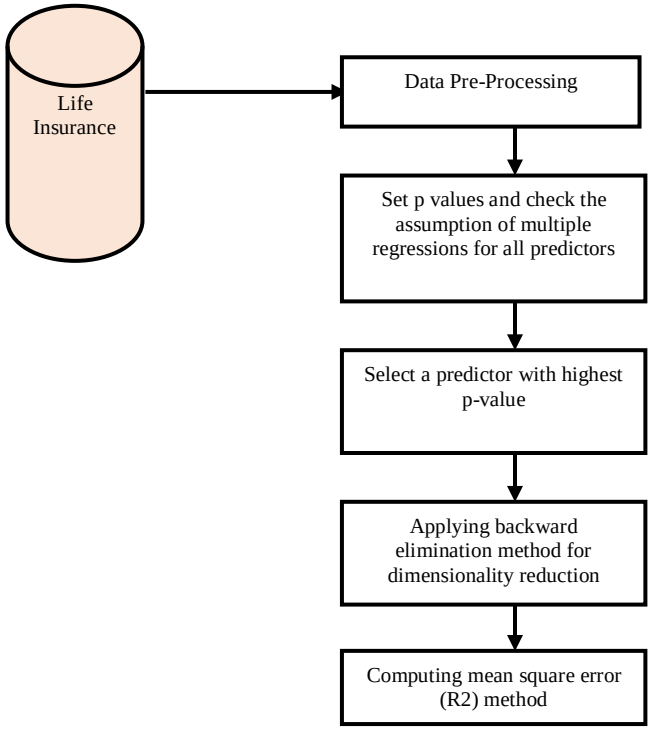

Fig 1: *Flowchart of proposed work*

Step 3: Fit the model with n predictors AND check p-value for every predictor:

- Use model to run the Multiple Linear Regression of the null hypothesis keeping the power of test close to one. This will ensure we will not accept the false negative results. Our Null hypothesis initially includes all predictors.
- These predictors decrease by one in every iteration. Thus, the Null hypothesis changes in every iteration and is equal to the difference between the total number of predictors and the total number of iteration.

Step 4: Select a predictor with the highest p-value and assign its p-value to p:

- The p-value of each predictor is calculated in step3. Out of the calculated p-values of predictors, the one with the highest p-value is used. It is denoted by p in step 3.
- If the p-value is too high, then it is not useful for machine learning models. Hence, the possibility of removing that predictor becomes high.
- If we remove all high p-value predictors in each iteration, then we will be left with predictors that may have a statistically significant strong correlation with the response variable.

Step 5 - Backward Elimination

- In this method, all features are selected to fit the model and the one, which is most statistically insignificant in the selected feature, is removed.
- In the next iteration, the model is fitted without this variable. This iterative process helps us to identify features that do not play an important role because they act as garbage within the data that increases time complexity of the machine learning models.
- In addition, they do not contribute to significantly increase in model accuracy as they fail to explain significant variance in the response variable.

After completing all the required steps, the last step is to compute the RMSE i.e. Root Mean Square error. This parameter basically is to compute the accuracy of any regression model.

3.2 Dataset

The dataset which is used in this work is taken from Prudential Life Insurance Assessment and this is available on kaggle. Kaggle is a online portal mainly built for competitions based on modeling and analytics. In this dataset, over hundreds of variables are used describing characteristics of life insurance applicants. The task was to predict the "Response" variable for each Id in the test set. FILE DESCRIPTIONS: We performed our analysis on a file named 'train.csv' in Prudential Life Insurance Assessment available on kaggel.com. It contains total 59382 rows and 128 columns. The 128 columns contain:
Id field, 26 features characteristic of policyholder and a response variable of 8 levels of ordered risk. There are total 60 categorical features and 13 continuous variables, 5 discrete variables and 48 dummy variables in this dataset. The 20% of the data in the train.csv file is used for testing and training was performed on the 80% of the data.

3.3 Tools Used

For python machine learning and data science, anaconda distribution provides more than 250 popular data science packages as well as the conda packages and virtual environment for Windows, MacOS and Linux. Conda makes easy and the quick run, install and upgrade complex machine learning and data science environment like Scikit-learn, keras, and TensorFlow. In Anaconda Repositories Python and R

conda packages are organised and compiled in a secure manner to get easy working of the optimized binaries. Anaconda made more than 1400 packages free for everyone in Anaconda Repository [11].

4 RESULT ANALYSIS

In this work python programming is used for preprocessing the dataset which is obtained from Prudential Life Insurance.

The attributes which are having more than 30% missing data those were discarded from the analysis process. Root Mean Square Error (RMSE) Parameter is used to shown the effectiveness of the classification algorithm [12]. RMSE is the measure that is used to show the difference between the sample value predicted by the applied model and the observed value. The main focus in this work is to minimize the error by applying the dimensionality reduction method. Even the work has shown that dimensionality reduction can be done with the help of feature selections and feature extraction methods like Principal Component Analysis [13], Correlation Based Feature Selection [14]. Various Machine learning algorithms are applied on this dataset including the proposed one i.e. Random Forest. Other than the proposed other algorithms which have been applied are Artificial Neural Network [15], Multiple Linear Regression Method [16] and Random Tree [17]. Results have been shown that Random forest classifier has demonstrated the least root mean square error with 0.384 estimation.

Comparison of Existing and Proposed

TABLE 1: *COMPARISON OF EXISTING AND PROPOSED ALGORITHM*

Algorithms	*Dimensionality reduction Method*	*RMSE*
Multiple Linear Regression	*PCA*	*2.0659*
Artificial Neural Network	*PCA*	*2.3369*
Random Tree	*PCA*	*2.9142*
Random Forest (Proposed)	*Backward Elimination*	*0.384*

In the above table we have shown how our proposed method that is backward elimination have improved performance over other method. The parameter taken for model evaluation is Random forest and the obtained RMSE (Root Mean Square Error) is 0.384.

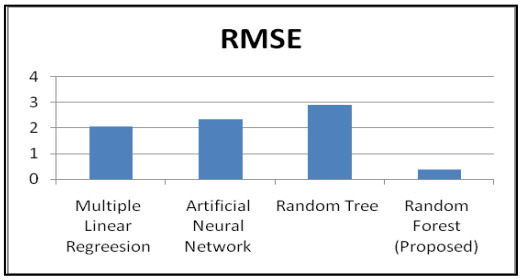

Fig 2: *Graph representation of classifier showing RMSE*

Graph representation of the previous table is shown which clearly indicates the improved result of our proposed method

TABLE 2: *COMPARISON OF EXISTING AND PROPOSED ALGORITHM REPRESENTING TRAINING ACCURACY*

Algorithms	*Dimensionality reduction Method*	*Accuracy (%)*
Naïve Bayes	*Backward Elimination*	*71.58*
Logistic Regression	*Backward Elimination*	*92.04*
Random Tree	*Backward Elimination*	*78.90*
Random Forest (Proposed)	*Backward Elimination*	*93.38*

Training accuracy of all the classifiers are shown in the given table which has been generated by applying the feature selection method i.e. Backward elimination and it is clearly observed that Random forest has achieved the highest training classifier as compared to others having accuracy of 93.38%.

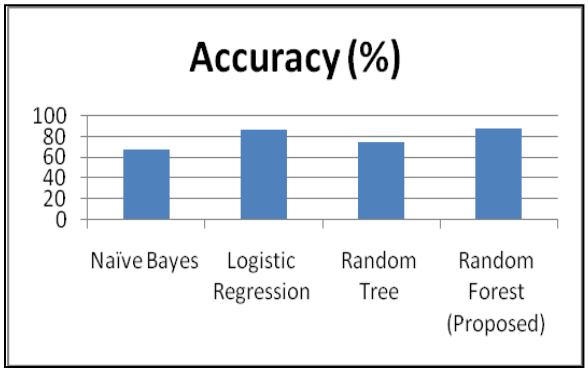

Fig 3: *Graph representation of classifier showing training accuracy*

TABLE 3: *COMPARISON OF EXISTING AND PROPOSED ALGORITHM REPRESENTING TESTING ACCURACY*

Algorithms	*Dimensionality reduction Method*	*Accuracy (%)*
Naïve Bayes	*Backward Elimination*	*67.60*
Logistic Regression	*Backward Elimination*	*86.93*
Random Tree	*Backward Elimination*	*74.51*
Random Forest (Proposed)	*Backward Elimination*	*88.19*

Testing accuracy of all the classifiers is shown in the table which are observed after implementing the feature selection method i.e. Backward elimination and it is clearly observed that Random forest has achieved the highest training classifier as compared to others having accuracy of 88.19%.

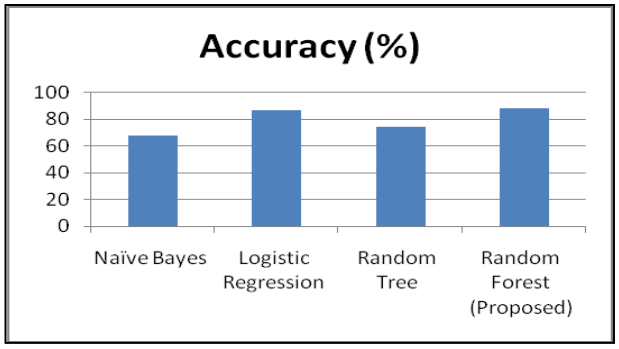

Fig 4: *Graph representation of classifier showing testing accuracy*

5 CONCLUSION AND FUTURE WORK

This exploration has explicit implication for the business condition. Data analytics is presently the pattern that is picking up criticalness among organizations around the world. In the life insurance domain, predictive modeling utilizing learning algorithms can give the eminent contrast in the manner which business is done as compared with the traditional strategies. Previously risk assessment, chance evaluation forever endorsing was led utilizing complex actuarial equations and more often than not was an extremely extensive procedure. Presently, with information expository arrangements, the work should be possible quicker and with better outcomes. Consequently, it would upgrade the business by enabling quicker administration to client, in this manner expanding fulfillment and unwaveringness. Future work identifies with the more top to bottom investigation of the issue and new techniques to manage specific systems. Client division is the division of the informational index into gatherings with comparable credits can be executed to section the candidates into gatherings with comparable qualities dependent on the characteristics present in the dataset. For instance, comparable work history, protection history and restorative history. Following the gathering of the candidates, prescient models can be executed to add to an alternate information digging approach for the disaster protection client informational collection.

6 .ACKNOWLEDGMENT

I would like to thank all the faculty members of Sagar Institute of Science & Technology.

7 REFERENCES

[1] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, “Critical analysis of Big Data challenges and analytical methods,” J. Bus. Res., 2017.
[2] Y. Joly et al., “Life insurance: Genomic stratification and risk classification,” Eur. J. Hum. Genet., 2014.
[3] Z. Ge, Z. Song, S. X. Ding, and B. Huang, “Data Mining and Analytics in the Process Industry: The Role of Machine Learning,” IEEE Access, 2017.
[4] A. C. Wuppermann, “Private Information in Life Insurance, Annuity, and Health Insurance Markets,” Scand. J. Econ., 2017.
[5] D. Hedengren and T. Stratmann, “Is there adverse selection in life insurance markets?,” Econ. Inq., 2016.
[6] J. M. Carson, C. M. Ellis, R. E. Hoyt, and K. Ostaszewski, “Sunk Costs and Screening: Two-Part Tariffs in Life Insurance,” J. Risk Insur., 2019.
[7] O. Devi, “Portfolio Rule - based Clustering at Automobile Insurance in Portugal PORTFOLIO RULE - BASED CLUSTERING AT AUTOMOBILE INSURANCE IN PORTUGAL.”
[8] N. Boodhun and M. Jayabalan, “Risk prediction in life insurance industry using supervised learning algorithms,” Complex Intell. Syst., 2018.
[9] C. Rubio-Bellido, A. Pérez-Fargallo, and J. Pulido-Arcas, “Multiple Linear Regressions,” 2018.
[10] R. Nair and A. Bhagat, “A Life Cycle on Processing Large Dataset - LCPL Rajit Nair,” vol. 179, no. 53, pp. 27–34, 2018.
[11] J. Phuong et al., “Automated retrieval, preprocessing, and visualization of gridded hydrometeorology data products for spatial-temporal exploratory analysis and intercomparison,” Environ. Model. Softw., 2019.
[12] T. Chai and R. R. Draxler, “Root mean square error (RMSE) or mean absolute error (MAE)? -Arguments against avoiding RMSE in the literature,” Geosci. Model Dev., 2014.
[13] D. J. Bartholomew, “Principal components analysis,” in International Encyclopedia of Education, 2010.
[14] M. Hall and L. a Smith, “Feature Selection for Machine Learning : Comparing a Correlation-based Filter Approach to the Wrapper CFS : Correlation-based Feature,” Int. FLAIRS Conf., 1999.
[15] M. Majumder, “Artificial Neural Network,” 2015.
[16] J. Fletcher, “Multiple linear regression,” BMJ, 2009.
[17] S. R. Safavian and D. Landgrebe, “A Survey of Decision Tree Classifier Methodology,” IEEE Trans. Syst. Man Cybern., 1991.