



REGRESSION BASED RISK ANALYSIS IN LIFE INSURANCE INDUSTRY

Fatma Bozyiğit¹ , Murat Şahin² , Tolga Gündüz³ , Cem Işık³ , Deniz Kılınç^{1*} 

¹İzmir Bakırçay University, Faculty of Engineering and Architecture, Department of Computer Engineering, İzmir, Turkey.

²Manisa Celal Bayar University, Faculty of Technology, Department of Software Engineering, Manisa, Turkey.

³Software Engineer, Compello, İstanbul, Turkey.

<https://doi.org/10.47933/ijeir.745343>

*Corresponding Author: deniz.kilinc@bakircay.edu.tr

(Received: 29.05.2020; Revised: 10.06.2020; Accepted: 15.06.2020)

ABSTRACT: Risk analysis is a crucial part for classifying applicants in life insurance business. Since the traditional underwriting strategies are time-consuming, recent works have focused on machine learning based methods to make the steps of underwriting more effective and strengthening the supervisory. The aim of this study is to evaluate the linear and non-linear regression-based models to determine the degree of risk. Therefore, four linear and non-linear regression algorithms are trained and evaluated on a life insurance dataset. The parameters of algorithms are optimized using Grid Search approach. The experimental results show that the non-linear regression models achieve more accurate predictions than linear regression models and the LGBM algorithm has the best performance among the all regression models with the highest R2, lowest MAE and RMSE values.

Keywords: Life insurance, Predictive analytics, Insurance analytics, Regression-based risk analysis.

1. INTRODUCTION

Insurance is one of the important business domains affected by digital transformation and technology [1]. Many data are produced in the insurance sector, such as requests from different channels, and sales from different platforms. When we consider the data types collected, it is possible to categorize types of insurances as follows i) Elementary insurances: They include standard data (model, color, engine etc.), past damage and repair information of cars for specific types such as motor insurance and traffic ii) Life insurances and private pension insurances: These types of insurances include detailed information from the demographic information of the person to the financial situation [2].

For a life insurance company, the traditional underwriting strategies are time-consuming and expensive. Therefore, finding ways to make the underwriting process faster and more cost efficient is crucial. Machine learning methods have proven effective in streamlining the method of underwriting and strengthening decision-making [3].

As stated in [4] underwriting process requires gathering detailed insurance claim records, which may be lengthy. The candidates are typically submitted to different medical examinations and the insurance provider must be supplied with all appropriate documentation. A research by [5] indicates that low capacity underwriting is a prominent operating concern among insurance firms surveyed in Bangladesh. Another risk to life insurance providers is not being prepared to

confront unfavorable competition. In this study it is aimed to apply predictive models to identify the degree of risks based on a dataset [6] containing 1,338 applications and to propose the most appropriate regression model for the risk management to optimize the underwriting process. The rest of the paper is organized as follows: Section 2 gives information about the materials and methods used. Section 3 presents evaluation results of the proposed study. Section 4 discusses and concludes the paper.

2. MATERIALS AND METHODS

2.1. Linear and Non-Linear Regression Methods

Regression which is one of the main areas of interest in statistical science, estimates the value of the dependent variable (y) based on the value of at least one independent variable (X) and explains the effects of changes in the independent variable on the dependent variable. In general regression models are examined in two groups as linear and non-linear.

Linear regression models use linear equations. A line is defined by a simple equation, measuring y from X , slope and intercept. The aim of linear regression is to find slope and intercept values which define the line that is closest to the samples data [7].

$$y(X) = \theta_0 + \theta_1 \cdot X_1 + \theta_2 \cdot X_2 + \dots + \varepsilon \quad (1)$$

Non-linear regression is more general than linear regression and can match the data with any model. It finds parameter values that generate the curve closest to the samples data [8]. Equation 1 and 2 show linear and non-linear regression formulas respectively. X_1 , X_2 and X_3 specifies independent variables, θ_0 is intercept, θ_1 , θ_2 and θ_3 are slope coefficients. ε is the random error term.

$$y = h_{\theta}(x) = \theta_0 + \theta_1 \cdot X_1 + \theta_2 \cdot X_2^2 + \theta_3 \cdot X_3^3 \dots + \varepsilon \quad (2)$$

2.2. Evaluated Regression Algorithms

PLS (Partial least squares) Regression is a statistical approach like the principal components' regression. The linear regression model is calculated by projecting the target variables and the observed variables into a new space [9].

RidgeRegression is also known as Tikhonov regularization, usually increases the performance of parameter identification problems giving precise approximate solutions in exchange for an acceptable level of bias [10].

LassoRegression is a method of linear regression that conducts both feature selection and regularization to improve the accuracy of estimation and generalizability of the mathematical model it generates [11].

LGBM (Light Gradient Boosting Method) is a gradient boosting approach based on tree learning algorithms. It supports parallelism and is designed to be capable of handling large-scale data [12].

RandomForest is an ensemble (collaborative) method that can execute both regression and classification tasks using multiple decision trees, and a method called Bootstrap Aggregation, widely known as bagging [13].

CART (Classification and Regression Trees) tree is a binary decision tree which is formed constantly by dividing a node into two child nodes, starting from the root node containing the entire training data set [14].

SVR (Support Vector Regression) is similar to the SVM (Support Vector Machine) classifier and characterized using kernels, sparse solution, and VC (Vapnik-Chervonenkis) control of the margin and the number of support vectors [15].

2.3. Dataset

In the study, “Insurance Premium Prediction” dataset containing 1,338 applications (observations) and 7 attributes (variables) is used [6]. This dataset contains 4 numerical attributes including age (in years), number of children, body mass index and medical costs. It also holds 3 nominal attributes: gender (male, female), smoking status (yes, no) and the region (southeast, southwest, northeast, northwest). The attribute named “expenses” is the dependent target variable and includes medical cost. The independent variables are the remaining 6 attributes. Table 1 shows the top 5 samples from the dataset.

Table 1. Top 5 samples of the dataset.

	age	sex	bmi	children	smoker	region	expenses
0	19	female	27.9	0	yes	southwest	16884.92
1	18	male	33.8	1	no	southeast	1725.55
2	28	male	33.0	3	no	southeast	4449.46
3	33	male	22.7	0	no	northwest	21984.47
4	32	male	28.9	0	no	northwest	3866.86

2.4. Pre-processing and Exploratory Data Analysis (EDA)

Pre-processing of data, also known as the data cleansing stage, includes eliminating noisy data, transforming variables, handling missing values, scaling the range of data, and label-encoding for categorical variables. Since the main purpose of the dataset is to predict medical expenses, the “expenses” variable is transformed to a “risk score” (1-100) using data transformation method named min-max scaler as the first step. Then, the exploratory data analysis (EDA) is applied to analyze the data based on various features such as age, body mass index, number of children, cigarette addiction and location against the current risk score.

EDA consists of univariate and multivariate analyzes. Thanks to EDA, the researchers are provided to understand the different distributions displayed by the features. Moreover, in the bivariate analysis, independent variables which are capable of impacting the target (dependent) variable can be analyzed. It can be seen intuitively which variables affect the target variable more strongly.

Although dozens of EDA analysis have been done in the study, the results of three EDA analyzes are presented and explained. Figure 1 that is a univariate analysis shows the distribution of the risk scores using a distribution plot having twenty bins. A significant part of the samples are in the range in which the risk score variable (target variable) takes a value

between 0-30. Besides there is also a remarkable distribution of samples in the 50-80 value range.

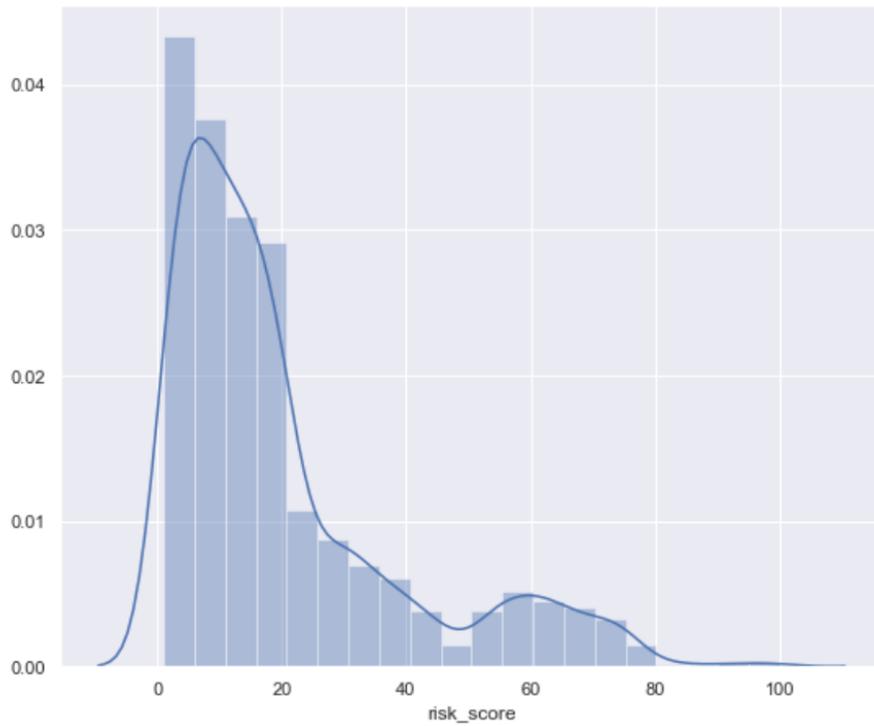


Figure 1. The distribution of risk scores.

Figure 2 presents the distribution of risk scores of smokers using a category plotter. It is observed that smokers have higher risk scores than non-smokers.

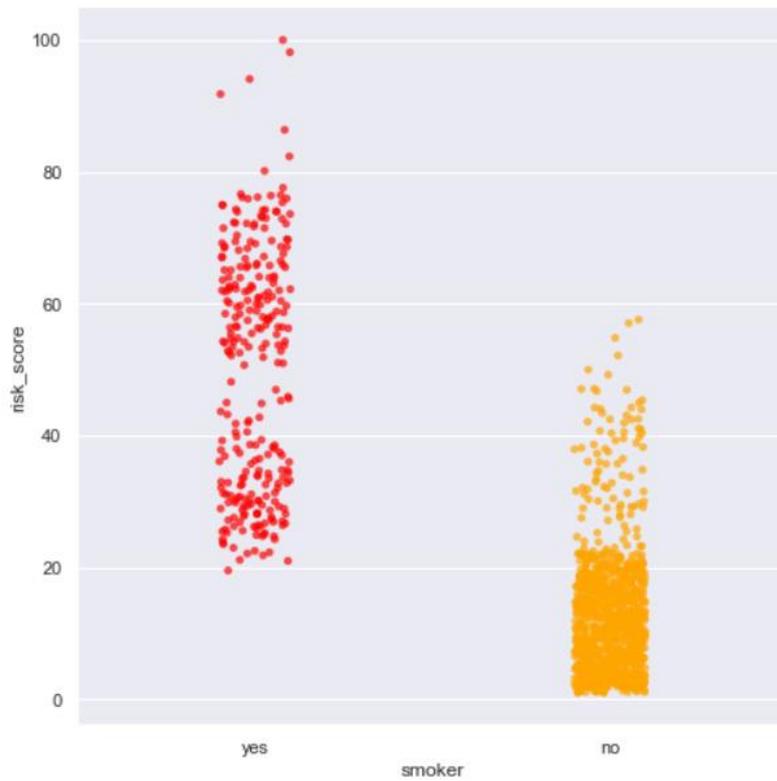


Figure 2. The distribution of smokers with their risk scores.

Figure 3 illustrates the relation between age and risk score considering smoking status. The risk scores of smokers are higher than the risk scores of non-smokers. Regardless of the smoking status, the risk scores of the people increased in accordance with their age, but it is still seen that the smokers' risk scores are higher than the non-smokers.

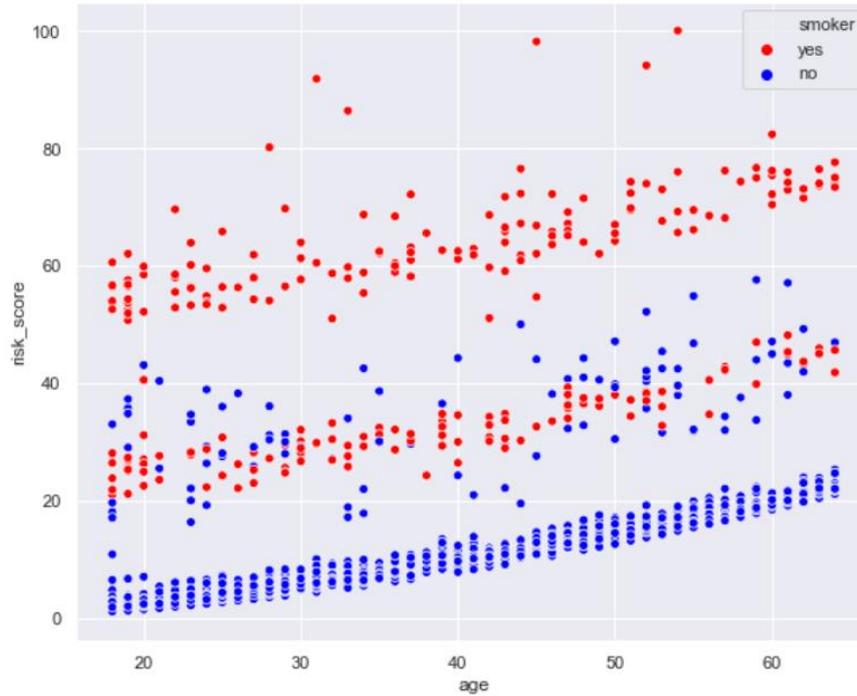


Figure 3. The relation between age and risk score considering smoking status.

3. EVALUATION METRICS

All trained regression models are evaluated utilizing on two common metrics named Root Mean Square Error (RMSE), and coefficient of determination (R^2) [16]. RMSE is a widely used measure of the difference between predicted and the real values of the model from the system that is trained. The RMSE values of a regression model is formulated as the square root of the mean squared error. R^2 indicates percentage variation of prediction values. The value of the R^2 is between 0 and 1. The formulas of metrics are defined as follows.

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (Y_{model,i} - Y_{obs,i})^2}{n}} \quad (3)$$

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_{obs,i} - Y_{model,i})^2}{\sum_{i=1}^n (Y_{obs,i} - \bar{Y}_{obs})^2} \quad (4)$$

where Y_{obs} are real values, and Y_{model} are the predicted values of model at position i .

4. EXPERIMENTAL RESULTS

In this work, four linear (LinearRegression, PLSRegression, RidgeRegression, LassoRegression) and four non-linear (LGBM, RandomForest, CART, SVR) regression algorithms are trained and obtained models are evaluated using Grid Search parameter optimizations (hyperparameter optimization) [17] which lets the selection of the best

parameters for an optimization problem from a list of parameters which is provided. The experimental results shown in Table 2 and Table 3 indicate that non-linear regression models perform more accurate predictions than linear regression models. CV_R^2 indicates the R^2 value obtained after performing cross validation.

Table 2. The evaluation performances of linear regression models.

	Model	R^2	CV_R^2	RMSE	MAE
1	LinearRegression	0.787472	0.732067	8.0	5.0
2	PLSRegression	0.766873	0.732054	9.0	6.0
3	RidgeRegression	0.767271	0.732054	9.0	7.0
4	LassoRegression	0.767237	0.732107	9.0	7.0

Findings yield as follows; Linear Regression algorithm has the highest performance with the lowest mean absolute error (MAE) value of 5.0 and lowest root-mean-squared error (RMSE) value of 8.0 between the linear regression models. On the other hand, LGBM shows the best performance between non-linear models with the highest R^2 , lowest MAE and RMSE values of 0.942658, 3.0 and 6.0, respectively, as compared to the other non-linear models.

Table 3. The evaluation performances of non-linear regression models.

	Model	R^2	CV_R^2	RMSE	MAE
1	LGBM	0.942658	0.912114	6.0	3.0
2	RandomForest	0.951977	0.907872	6.0	3.0
3	CART	0.903022	0.885989	7.0	4.0
4	SVR	0.893027	0.827012	7.0	4.0

After hyperparameter optimization, obtained optimum parameters' values of the best performing algorithm LGBM is shown in Table 4.

Table 4. The optimum parameters' values of LGBM regression model.

Parameter Name	Value
colsample_bytree	0.9
learning_rate	0.5
max_depth	2
n_estimators	20

5. CONCLUSIONS

With the recent advances of technology, data analytics has become an important trend. In the area of life insurance, predictive analytics using learning algorithms have made a significant difference in how business is conducted comparing to traditional approaches. In this research, four linear algorithms named LinearRegression, PLSRegression, RidgeRegression, LassoRegression and four non-linear regression algorithms named LGBM, RandomForest, CART, SVR are implemented, trained and tested on a publicly available insurance dataset. Hyperparameter optimization is performed to find the best parameters of the algorithms. The experimental results show that non-linear models have better accuracy than linear ones. Finally, it can be concluded that tree-based non-linear regression models (LGBM, RandomForest) are promising to forecast the risk score of applicants and can be used in real-life scenarios.

ACKNOWLEDGMENTS

Funding for this work was partially supported by the Research and Development Center of Compello accredited on Turkey - Ministry of Science.

REFERENCES

- [1] Burri, R. D., Burri, R., Bojja, R. R., & Buruga, S. (2019). Insurance Claim Analysis using Machine Learning Algorithms. *International Journal of Advanced Science and Technology*, 127(1), 147-155.
- [2] Boodhun, N., & Jayabalan, M. (2018). Risk prediction in life insurance industry using supervised learning algorithms. *Complex & Intelligent Systems*, 4(2), 145-154.
- [3] Bhalla, A. (2012). Enhancement in predictive model for insurance underwriting. *Int J Comput Sci Eng Technol*, 3, 160-165.
- [4] Wuppermann, A. C. (2017). Private Information in Life Insurance, Annuity, and Health Insurance Markets. *The Scandinavian Journal of Economics*, 119(4), 855-881.
- [5] Mamun, D. M. Z., Ali, K., Bhuiyan, P., Khan, S., Hossain, S., Ibrahim, M., & Huda, K. (2016). Problems and prospects of insurance business in Bangladesh from the companies' perspective. *Insur J Bangladesh Insurance Acad*, 62, 5-164.
- [6] Web Access (January 2020), <https://www.kaggle.com/noordeen/insurance-premium-prediction>
- [7] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (Vol. 821). John Wiley & Sons.
- [8] Gallant, A. R. (2009). *Nonlinear statistical models* (Vol. 310). John Wiley & Sons.
- [9] Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley interdisciplinary reviews: computational statistics*, 2(1), 97-106.
- [10] Saleh, A. M. E., Arashi, M., & Kibria, B. G. (2019). *Theory of Ridge Regression Estimation with Applications* (Vol. 285). John Wiley & Sons.
- [11] Reid, S., Tibshirani, R., & Friedman, J. (2016). A study of error variance estimation in lasso regression. *Statistica Sinica*, 35-67.
- [12] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in neural information processing systems* (pp. 3146-3154).
- [13] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- [14] Steinberg, D. (2009). CART: classification and regression trees. In *The top ten algorithms in data mining* (pp. 193-216). Chapman and Hall/CRC.
- [15] Awad, M., & Khanna, R. (2015). Support vector regression. In *Efficient Learning Machines* (pp. 67-80). Apress, Berkeley, CA.
- [16] Basaran, K., Özçift, A., & Kılınç, D. (2019). A new approach for prediction of solar radiation with using ensemble learning algorithm. *Arabian Journal for Science and Engineering*, 44(8), 7159-7171.
- [17] Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of machine learning research*, 13(Feb), 281-305.