Advances in Parallel Computing Algorithms, Tools and Paradigms
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC220027

A Light Gradient Boosting Machine Regression Model for Prediction of Agriculture Insurance Cost over Linear Regression

Purna Syam Chand S[a,1] and G.Divya[b]
[a,b] *Saveetha School of Engineering,*
Saveetha Institute of Medical and Technical Sciences,
Saveetha University, Chennai, Tamilnadu, India. Pincode:602105

Abstract. To increase accuracy for the prediction of agriculture insurance claim cost based on crop insurance data.Gradient Boosting Machine (LGBM) and linear regression models are tested with total Samples 6022 for n=7 iterations to predict accuracy. LGBM works based on decision tree algorithm and linear based on fitted regression equation. **:**The coefficient of determination values of proposed LGBM regression (92.52%) and linear regression (72.47%) are obtained. There was a statistical significance between LGBM regression and linear regression (p=0.001).Prediction of agriculture insurance claim cost LGBM regression technique produces significantly better performance than the linear regression technique.

Keywords. Agriculture Insurance, Novel Light Gradient Boosting Machine Regression, Linear Regression, Data Science, Machine Learning

1. Introduction

Insurance is to provide financial protection(Mahdzan and Diacon 2008)[1]. This study describes non-life insurance, particularly agriculture insurance. Prediction of agriculture claim cost is that the data analysis of estimating the premium of agriculture supported some influence factors which are presented within the dataset.Estimating the claim cost will be more useful to the insurance companies ("Insurance Claim Analysis Using Machine Learning Algorithms" 2019)[2]. It will allow the optimization of insurance pricing and reduction of fraudulent claims (Belhadji, Dionne, and Tarkhani 2000)[3]. [4]Its estimates of agriculture claim cost are often used for other applications like risk assessment within the agriculture insurance industries, and it makes agriculture coverage cheaper for people (Grace and Klein 2003).

Agriculture Insurance claim cost prediction is carried out by researchers and 4 related research articles in IEEE Digital Xplore and 12 articles are published in the google scholar.[5](Pesantez-Narvaez, Guillen, and Alcañiz 2019) represents the positive outcome of analyzing of computerized information to estimate insurance

[1] Purna Syam Chand S, Corresponding author.

premiums for motor vehicles(Hultkrantz, Nilsson, and Arvidsson 2012)[6], speaks about the significance of PAVD insurance plans and they designed personalize premium calculator for the insurance cimpanies.[7](Fang, Jiang, and Song 2016) proposed a regression model to forecast insurance customer profitability with different algorithms like RF,LR,SVM based on R-squared value Random Forest(RF) got better performance.[8](Hanafy and Ming 2021) proposed with various machine learning approaches to predict auto insurance from that also random forest regression model was performed well. Comparing all the surveyed articles[5](Pesantez-Narvaez, Guillen, and Alcañiz 2019), research work seems to be base paper which gives estimation of vehicle insurance claim cost with various regression techniques[9] (Coble et al. 2003).

Previously, there are several technique are applied for prediction of agriculture cost.s.[10,11] (Kumar et al. 2006; Danda et al. 2010; Gopalakannan, Senthilvelan, and Ranganathan 2012) . Now the emerging trend in this topic motivates to carry out this this work.Based on the literature survey, the predicted models appear to perform with less accuracy. The aim of the study is to implement a better regression model for predicting the agriculture insurance claim cost.

2. Materials and Method

The study setting of the proposed work is done in Saveetha University. The number of groups identified for this study is 2.

Table 1: Dataset Information

Attribute	Description of attribute
Area-Type	0-Small 1-Medium 2-Big
Education	1- Secondary 2- Higher secondary 3- Junior college
Employment status	1-Employed 2-Unemployed
Gender	0- Male 1-Female
Income	Income in Dollars($)
Marital Status	0-Single 1-Married
Months Since Last Claim No.of open complaints No.of Policies	Last claimed month Count of crop related complaints Policies taken by the user
Rainfallen	0-No 1-Yes
Policy Type	0-Food Security 1-Poverty Reduction
Claim Reason	0-Diseases 1-Drought 2-Floods 3-Land Fertility
Crop Size	0-Small 1-Medium 2-Big
Claim Amount(Dependent Variable)	Claimed Amount by the person in ($'s)

The group 1 is a Novel LGBM (Light gradient boosting machine) regression and group 2 is Linear regression of Machine Learning Techniques. Using clinical analysis[12] (Bärtl and Krummaker 2020) 95% confidence and pretest power 80% was carried out 20 samples In this paper, the dataset is taken from the[13] ("Kaggle: Your Machine Learning and Data Science Community".) which is agriculture insurance data.The taken dataset has various attributes and it has 14 variables as shown in Table 1.

2.1 LGBM Regression

It is a distributed and fast, scalable gradient boosting architecture supported decision tree algorithm in Data Science. It can be applied for regression, classification and ranking and traditional approaches of machine learning, because it is based on a decision tree algorithm. LGBM algorithm splits the tree in the order of leaf wise with the simplest fit, other boosting algorithms it splits the tree in the order of level wise or depth wise. leaf- wise algorithm can reduce loss than level-wise algorithm. The pseudo code of LGBM regression shown in Table 2.

Table 2: Pseudo code for LGBM Regression

Input • Import the dataset and required packages
1. Preprocess the data which is removal of unnecessary data. //Data Preprocessing
0. Identification of Dependent and Independent Variables. Setting up the data for LGBM regression. //Initialization
0. Training the model • Using LightGBM Classifier • Define a lgbmregressor() function • Use lgb.fit() to fit the model //Generating the model
0. Testing the model • Split the dataset into two parts Training(80%),Testing(20%)
0. Evaluating the model • Print the regression equation Output: R-Squared values,MAE,MSE,RMSE.

2.2 Linear Regression

Multiple linear regression is a common predictive analysis to forecast the prediction. It is used to establish the relationship between dependent and independent variables. Categorical data or continuous data cam be independent variable. The relationship among the variables can be observed through the regression equation mentioned in equation(1).The pseudo code for linear regression shown in Table 3.

Table 3: Pseudo code for Linear Regression

Input • Import the dataset and required packages
1. Preprocess the data which is removal of unnecessary data.
2. Describe the Dependent and Independent Variables. //Initialization
3. Training the model • Define a LinearRegression() function • Use linearreg.fit() to fit the model between x_train and y_train.
4. Testing the model • Split the dataset into two parts Training(80%),Testing(20%)
5. Evaluating the model • Print the regression equation Output: R-Squared values,MAE,MSE,RMSE.

$$Y=b+ 1X_1+ {}_2X_2+\ldots+ {}_iX_i \qquad (1)$$

Where Y : Dependent Variable
b : Intercept
$_i$: Slope for X_i
X : Independent Variable

For comparing both the models,the dataset has been trained with seven different sample sizes, the R- squared values are recorded. The system configuration is used for the algorithm to run in a 64-bit Operating System,4GB RAM PC and used Windows 10,Python3,Jupyter Notebook for software specification. To estimate the performance of the training model, the data has been splitted for training and testing to validate the dataset. The model will be evaluated with the fit() function which has the metrics function to validate the model of R- Square values and Errors (MAE, MSE, RMSE).

The Dependent variable is the claim amount and some of the independent variables are Area type, Education, Employment status, Gender, Income, Marital Status, Months since last claim, No. of open complaints, No. of policies, Rainfallen, policy type, claim reason, crop size. to evaluate the performance of the algorithm Independent sample t-Test is carried out to measure the performance analysis.

2.3 Statistical Analysis

Statistical software tool used in this work is tIBM SPSS version 21. Independent Sample t-test was performed to compare the performance of Novel LGBM and Linear regression for the prediction of better accuracy.

3. Results

The dataset is splitted into different sizes to measure the r-squared values (accuracy). The performance metrics of lgbm regression are represented in Table 4 and performance metrics of linear regression are shown in Table 5. SPSS is used for statistical analysis. In SPSS, a dataset is prepared using 7 iterations for lgbm regression and 7 iterations for linear regression. GroupId is labeled as grouping variable and Accuracy and Loss is given as the testing variable. GroupId is given as 1 for Novel Light Gradient Boosting Machine regression and 2 for Linear regression. Group statistics using SPSS is analyzed on given dataset and shown in Table 6. The Group statistics consists of mean and standard deviation and standard error mean of two groups.

Table 4: Seven iterations for Performance metrics of lgbm regression model(92.52%) for Sample Size=6022

Iterations (n)	R-Squared values(accuracy) in percentage %	Loss(in percentage %)	MAE(Mean absolute error)	MSE(Mean squared error)	RMSE(Root mean squared error)
1	92.42	7.58	0.003642	0.000030	0.005464
2	93.22	6.78	0.003421	0.000027	0.005157
3	93.05	6.95	0.003349	0.000026	0.005109
4	92.09	7.91	0.003222	0.000024	0.004939
5	92.09	7.91	0.003007	0.000021	0.004624
6	91.93	8.07	0.002930	0.000020	0.004482
7	92.87	7.13	0.002841	0.000019	0.004369

Table 5: Seven iterations for performance metrics of linear regression(72.01%)for sample size=6022

Iterations (n)	R-Squared values(accuracy)in percentage %	Loss(in percentage %)	MAE(Mean absolute error)	MSE(Mean squared error)	RMSE(Root mean squared error)
1	72.63	27.37	0.006125	6.883402	0.008296
2	72.49	27.51	0.005789	6.136278	0.007833
3	72.22	27.78	0.005577	5.732259	0.007571
4	71.63	28.37	0.005353	5.291937	0.007274
5	71.60	28.40	0.005113	4.834342	0.006952
6	71.55	28.45	0.004942	4.489773	0.006700
7	72.00	28.00	0.004793	4.247547	0.006517

Table 6: Group Statistics T-test for LGBM Standard Error Mean (0.19647) and Linear (0.16781)

Groups		**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
Accuracy	**LGBM**	7	92.5243	.51980	.19647
	Linear	7	72.0171	.44399	.16781
Loss	**LGBM**	7	7.4757	.51980	.19647
	Linear	7	27.9829	.44399	.16781

In this study, from the observation lgbm regression technique acheived accuracy of 92.52% than the linear regression and Novel Light Gradient Boosting Machine regression got better significant results while compared with the linear regression technique

In the SPSS,confidence interval at 95% and level of significance as 0.005 by Independent T-test on the dataset as shown in Table 7. LGBM regression and linear regression algorithms are significantly different to each other. Then a simple line graph is plotted using GroupId as X-axis and mean of accuracy and loss as Y-Axis then displaying the accuracy and loss of lgbm regression and linear regression are shown in Figure. 1.

Table 7: Independent Sample T- Test is applied for the dataset fixing the confidence interval as 95 % and level of significance as 0.05.

		Levene's test for equality of variances		**T-test for equality means**						
									95% confidence interval	
		f	Sig.	t	df	Sig. (2-tailed)	Mean difference	Std.Error difference	Lower	Upper
Accuracy	Equal variances assumed	.589	.458	79.368	12	.001	20.50714	.25838	19.9448	21.07010
	Equal Variances not assumed			79.368	11.714	.001	20.50714	.25838	19.94265	21.07613
Loss	Equal variances assumed	.589	.458	-79.368	12	.001	-20.50714	.25838	-21.07010	-19.94418
	Equal variances not assumed			-79.368	11.714	.001	-20.50714	.25838	-21.07163	-19.94265

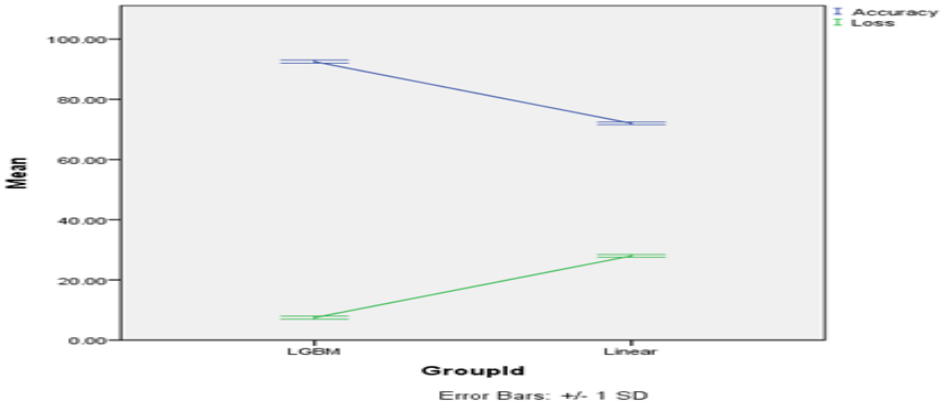

Fig. 1. Comparison of lgbm regression and linear regression in measure of accuracy(92.52) and mean loss(72.01). The mean accuracy of lgbm regression is significantly better than the linear regression. X Axis: LGBM regression vs Linear regression, Y Axis:Mean of Accuracy and Mean of loss

The Bar graph Fig. 2 shows the comparison of mean absolute error(MAE), mean squared error(MSE),root mean squared error(RMSE) of Novel Light Gradient Boosting Machine regression and linear regression. LGBM regression gives more accuracy. LGBM regression gives very less error compared with linear regression.

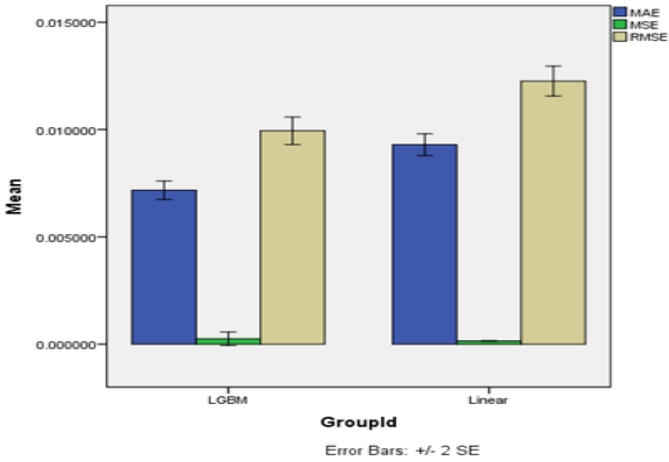

Fig. 2. Comparison of lgbm regression and linear regression in terms of MAE, MSE, RMSE. X Axis:LGBM vs Linear,Y Axis:Mean of MAE,MSE,RMSE

4. Discussion

In this study, the lgbm regression has improved r-squared value than the linear regression ($p<0.001$,Independent Sample T Test). The mean r-squared value for prediction of agriculture insurance using lgbm regression (mean accuracy=92.52,mean loss=7.48) than linear regression (mean accuracy=72.01,mean loss=27.99).The bar

graph Fig. 2, represents the MAE, MSE, RMSE metrics for LGBM and Linear Regression.

The similar findings of the related work found in the previous study are discussed. [14](Subudhi and Panigrahi 2020) proposed classification model for predicting insurance fraud using the Decision Tree and Support Vector Machine algorithms from that based on accuracy SVM algorithm (60.31%) got the best model.[7] (Fang, Jiang, and Song 2016) proposed a regression model to forecast insurance customer profitability with different algorithms like RF, LR, SVM based on R-squared value Random Forest (RF) got better performance. [5](Pesantez-Narvaez, Guillen, and Alcañiz 2019) proposed a regression models to predict insurance premium using Computerized data, XG Boost and logistic regression models implemented from that XGBoost(Linear Booster) regression model got 65% accuracy. [15](Sukono et al. 2018) discussed the model which estimate vehicle insurance premium using a Bayesian method with 70% accuracy. [8](Hanafy and Ming 2021) they implemented different types of machine learning techniques for auto insurance data,at last they analyzed RF to get better accuracy but they got a high error rate which is (0.1323).

Our institution is passionate about high-quality evidence-based research and has excelled in various fields. We hope this study adds to this rich legacy.

In this work,LGBM regression gives better performance with very less error compared with previous algorithms . The limitation of the proposed method is overfitting of the model and still the model has to be trained to predict the least loss error. In future scope, More number of parameters like type of pests and type of diseases can be included to test the model for better prediction.

5. Conclusion

From the results and in comparison, that lgbm regression technique appears to be a better model(92.52%) over Linear Regression (72.01%) for Prediction of agriculture insurance premium.

References

[1] Mahdzan, Nurul Shahnaz, and Stephen Diacon. 2008. “Protection Insurance and Financial Wellbeing.” UK Financial Services Research Forum. https://doi.org/10.13140/RG.2.1.3876.3046.

[2] Insurance Claim Analysis Using Machine Learning Algorithms, 2019. International Journal of Innovative Technology and Exploring Engineering. https://doi.org/10.35940/ijitee.f1118.0486s419.

[3] Belhadji, El Bachir, George Dionne, and Faouzi Tarkhani. 2000. “A Model for the Detection of Insurance Fraud.” Geneva Papers on Risk and Insurance - Issues and Practice. https://doi.org/10.1111/1468-0440.00080

[4] Grace, Martin F., and Robert W. Klein. 2003. “Homeowners Insurance: Market Trends, Issues and Problems.” SSRN Electronic Journal. https://doi.org/10.2139/ssrn.816927

[5] Pesantez-Narvaez, Jessica, Montserrat Guillen, and Manuela Alcañiz. 2019. “Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression.” Risks. https://doi.org/10.3390/risks7020070.

[6] Hultkrantz, Lars, Jan-Eric Nilsson, and Sara Arvidsson. 2012. “Voluntary Internalization of Speeding Externalities with Vehicle Insurance.” Transportation Research Part A: Policy and Practice. https://doi.org/10.1016/j.tra.2012.02.011.

[7] Fang, Kuangnan, Yefei Jiang, and Malin Song. 2016. “Customer Profitability Forecasting Using Big Data Analytics: A Case Study of the Insurance Industry.” Computers & Industrial Engineering. https://doi.org/10.1016/j.cie.2016.09.011.

[8] Hanafy, Mohamed, and Ruixing Ming. 2021. "Machine Learning Approaches for Auto Insurance Big Data." Risks. https://doi.org/10.3390/risks9020042.

[9] Coble, Keith H., Terry Hanson, J. Corey Miller, and Saleem Shaik. 2003. "Agricultural Insurance as an Environmental Policy Tool." Journal of Agricultural and Applied Economics. https://doi.org/10.1017/s1074070800021350.

[10] Danda, Anil Kumar, M. R. Muthusekhar, Vinod Narayanan, Mirza F. Baig, and Avinash Siddareddi. 2010. "Open versus Closed Treatment of Unilateral Subcondylar and Condylar Neck Fractures: A Prospective, Randomized Clinical Study." Journal of Oral and Maxillofacial Surgery: Official Journal of the American Association of Oral and Maxillofacial Surgeons 68 (6): 1238–41.

[11] Gopalakannan, S., T. Senthilvelan, and S. Ranganathan. 2012. "Modeling and Optimization of EDM Process Parameters on Machining of Al 7075-B4C MMC Using RSM." Procedia Engineering 38: 685–90.

[12] Bärtl, Mathias, and Simone Krummaker. 2020. "Prediction of Claims in Export Credit Finance: A Comparison of Four Machine Learning Techniques." Risks. https://doi.org/10.3390/risks8010022.

[13] Kaggle: Your Machine Learning and Data Science Community, n.d. Accessed March 30, 2021. https://www.kaggle.com/.

[14] Subudhi, Sharmila, and Suvasini Panigrahi. 2020. "Use of Optimized Fuzzy C-Means Clustering and Supervised Classifiers for Automobile Insurance Fraud Detection." Journal of King Saud University - Computer and Information Sciences. https://doi.org/10.1016/j.jksuci.2017.09.010.

[15] Sukono, Sukono, Riaman, E. Lesmana, R. Wulandari, H. Napitupulu, and S. Supian. 2018. "Model Estimation of Claim Risk and Premium for Motor Vehicle Insurance by Using Bayesian Method." IOP Conference Series: Materials Science and Engineering. https://doi.org/10.1088/1757-899x/300/1/012027.