e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:07/July-2022 Impact Factor- 6.752 www.irjmets.com

A STUDY ON CAR INSURANCE PURCHASE PREDICTION USING MACHINE LEARNING

Shwetha Shree K*1, Mohan Murali P*2

*1Master Of Computer Application, East West Institute Of Technology(VTU) ,
Bangalore, India.

ABSTRACT

With the continued advancement of AI, projects utilising AI strategies to exploit potential data has grown to be a hotly debated subject in the evaluation of significant insurance companies. The primary highlights impacting auto recharge are mined in this article, which dissects the pieces of accident coverage information. We examine lifting machine calculation, slope lifting tree, and irregular timberland (RF) (LightGBM). The LightGBM model has the best prevalence and heartiness, according to the testing data. Elements such as the vehicle protection business channel, NCD, vehicle age, and the cost of acquiring a new vehicle have a stronger impact on whether or not to recharge protection.

Keywords: Car Protection, Feature Designing, LightGBM, Data Investigation.

I. INTRODUCTION

As the number of cars on the road grows, corporations will place a greater emphasis on precision marketing. The competitiveness of large insurance organisations has relied on extracting critical knowledge and information from users, goods and services from massive customer data, and the acquisition of more customer resources. One method to gain a competitive edge is to use machine learning and data mining to improve goods and services. [1]. One of the most prevalent strategies in data preprocessing is feature selection. It focuses on removing superfluous or redundant features from the first information and picking a small number of key features as a dimension reduction method [2Suyeon Kang et al. [3] suggested a new component choice calculation for accumulated information inquiry that has exceptional adaptability be applied to actual accident protection data, and it deals with the problem of informational collecting that is difficult to demonstrate. Alshamsi and others [4] use irregular backwoods computations to help guarantors predict client decisions in order to provide more ruthless sorts of assistance. In comparison to information handling and angle lifting tree calculation, the LightGBM calculation has evident advantages [5]. Comparing several item expectation computations, Yanmei Jiang et al. [6] discovered that the LightGBM model had the best presentation. This article examines over 60,000 accident coverage records and employs the LightGBM calculation model to identify the key characteristics that influence customer reestablished protection, allowing organisations to more effectively nurture advertising tactics.

II. INTERPRETING INFORMATION AND ENGINEERING FEATURES

1 Cleaning up data and renaming features

We analyse the data based on current business facts to comprehend the effects that each component provides. The information is then preprocessed. The main tasks are: removing inaccurate data, filling in lacking worth, reducing include aspect, and so on. There are 28 element factors and 65,535 total raw data points used in this article. The following are the characteristics of the client collision protection information: Whether the region tag; use property; vehicle type; vehicle reason; new vehicle acquisition cost; vehicle age; protection type; NCD; risk class (A base, E most elevated); arrangement number; begin date ; final day; the business channel for auto insurance; brand of vehicles; automobile series; security element; Year of reestablishment; categorization for protection; If the region tag; apply property; vehicle class; vehicle cause; cost of purchasing a new vehicle; vehicle age; protection type; NCD; risk class (A base, E most elevated);client classification; the protected individual's orientation; the protected individual's age; whether the vehicle is protected from harm; whether the vehicle is protected from burglary; whether the car is protected from guaranteed people; The amount of protection; the amount of the marking instalment; the number of cases; Amount of money that has been agreed upon. Following a thorough examination of the data, It turns out that the approach number, start date, end date, vehicle brand, and vehicle series have no impact on whether the protection is restored. These components that

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:07/July-2022 Impact Factor- 6.752 www.irjmets.com

are repeated are easily removed. We also omit the protection property and restoration year since they are overly applicable to reestablishment or not. We rename the elements as shown in Table to make it simpler to work with the work in the future.

Table 1. Features and field name

FieldName	DisplayName	Field Type	MaxMask	Feature Class Field Name	Field Lookup	Lookup Lucity ID
PA_ADR_STR	Street Name	String		ADDRESS		
PA_ADR_TY	Street Type	String	4x			
PA_AREA	Area	Double	-nnnnnnnnn...			
PA_BR_CD	Default WO Cat	String	10x			
PA_CITY_CD	City	Short	nnnn			
PA_COUN_CD	County	Short	nnnn			
PA_DIST_CD	District	Short	nnnn			
PA_GPS	GPS Flag	Boolean				
PA_ID	Plant Rec #	Long	nnnnnnnn	LUCITYID		
PA_LOCATION	Location	String	100x			
PA_MLOCAT	Map Location	String	30x			
PA_NAME	Plant Name	String	40x	NAME		
PA_NOWORK	No WO/PM/Req	Boolean				
PA_NUMBER	Plant ID	String	20x	FACILITYID		
PA_OPENDT	Date Opened	Date	mm/dd/yyyy			
PA_OWN_CD	Owner	Short	nnnn			
PA_POSTAL	Zip	String	15x			
PA_PROPTAG	Property ID Tag	String	52x			

2 Eigenvalue and Missing Value Filling

Handling There are several gaps in the information's highlights, which are filled in as follows: 1Since each vehicle type and vehicle proposal have one incentive missing, we can easily eliminate these two details. 2 NCD contains 11 lacking characteristics, and We only remove the sample data for the values that are missing. ③ Almost 50,000 attributes are missing from the gamble class. Because this element can have a bigger effect on the model's results, we choose 0 to fill in the missing data. ④The protected person's orientation esteem is likely to be lacking in the thousands. Fill male or female having a 50% chance of being either male or female. Many highlights are text and cannot be incorporated into the model.A few eigenvalues should be determined, and the matrix should be divided into a few spans. Table 2 shows the tasks that are explicitly stated.

Table 2. Segmentation and quantification of eigenvalues

Total Variance Explained

Factor	Total Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	6.249	52.076	52.076	5.857	48.758	48.758	2.950	24.583	24.583
2	1.229	10.246	62.322	.808	6.719	55.478	2.855	22.927	47.510
3	.719	5.992	68.313	.360	3.000	58.478	1.412	11.769	59.279
4	.673	5.609	73.923						
5	.561	4.678	78.600						
6	.503	4.192	82.791						
7	.471	3.927	86.718						
8	.389	3.240	89.958						
9	.368	3.066	93.024						
10	.320	2.675	95.699						
11	.301	2.515	98.214						
12	.215	1.786	100.000						

Extraction Method: Principal Axis Factoring.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:07/July-2022 Impact Factor- 6.752 www.irjmets.com

III. INDEX FOR MODEL PERFORMANCE EVALUATION

A model order execution assessment list is created using the characterisation precision rate and every class is assumed to have a similar commitment to the exactness rate. This paper has a 5:1 ratio between class 0 (no) and class 1 (yes), which indicates some degree of lopsidedness. In order to examine the model's execution of characterisation, assessment indicators including the positive class review rate, F1 esteem, and AUC esteem are utilised. Table 3 shows the disarray network in the parallel characterization problem.

Table 3. Confusion matrix

Confusion Matrix

	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TPs)	False Positives (FPs)
Predicted Negative (0)	False Negatives (FNs)	True Negatives (TNs)

IV. BUILDING A MODEL

Small batches are usually used to train machine learning algorithms, with no memory constraints on the quantity of the training data. At each iteration, The entire training set must be iterated through using the GBDT algorithm. The main goal of LightGBM is to solve the problems GBDT has with handling big volumes of data. A decision tree-based learning system called LightGBM uses gradient boosting to deliver efficient parallel training at faster training speeds, reduced memory requirements, greater accuracy, and speedier processing of data. A training set has been created using the expertise of the company and the data has been turned into training samples that the model can recognise.The overall process of the model is shown in Fig.1.

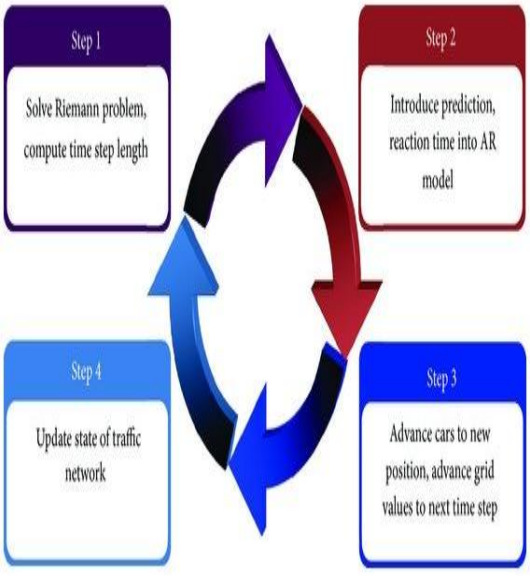


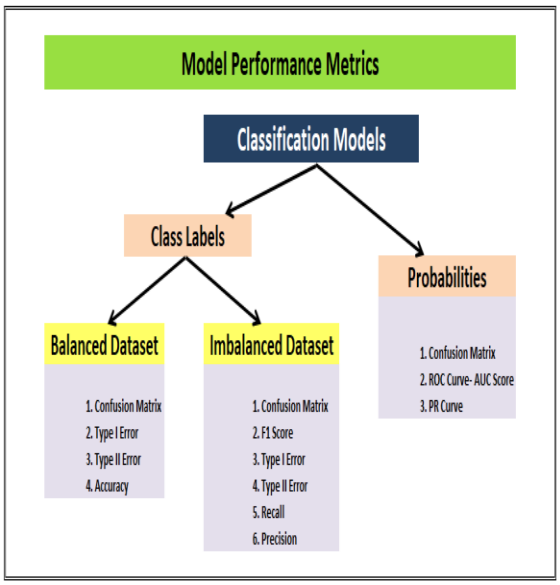
Figure 1. The model's overall methodology

V. RESULTS

As shown in Table 4, Compare the handled informative sets to the RF and GBDT calculation models after organising them.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:07/July-2022 Impact Factor- 6.752 www.irjmets.com

Table 4. Execution using different models



The information in Table 4 shows that the LightGBM calculation model has a specific amount of advancement in the correlation of these three assessment markers, aside from the F1 esteem, which is marginally less than the RF estimate. Figure 2 depicts the ROC bend. The LightGBM calculation, on average, has a higher order influence. As shown in Fig.3, the components impacting vehicle protection restoration are grouped by relevance in the LightGBM calculation testing. The elements that determine vehicle reestablishment are primarily the channel for business protecting vehicles, NCD, new vehicle acquisition cost, and age, as shown in the graph.As a result of this finding, insurance companies might use more targeted marketing strategies to increase their income.

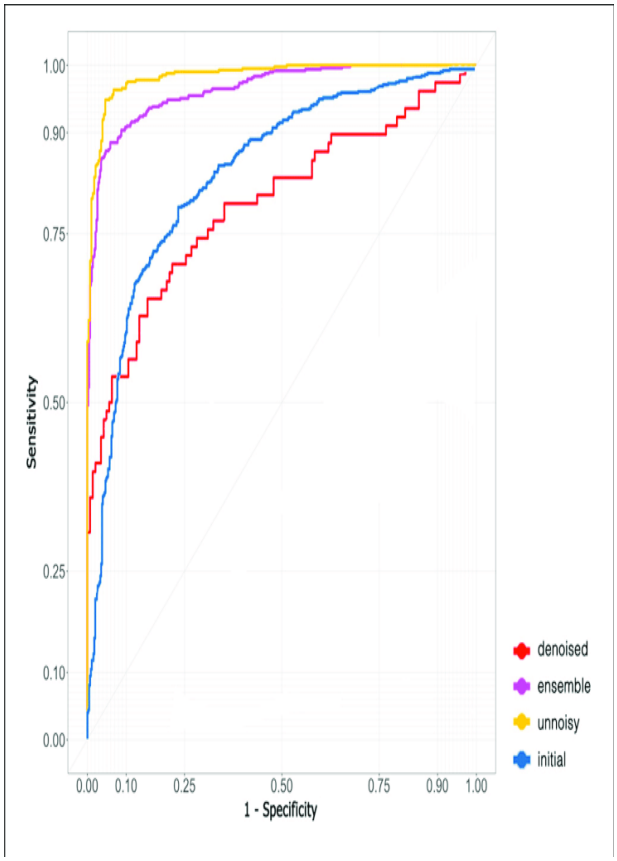


Figure 2. ROC traces for several models

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:04/Issue:07/July-2022 Impact Factor- 6.752 www.irjmets.com

Figure 3. Features that affect renewal in order of relevance

VI. CONCLUSION

In this research, we used the Random Forest Classifier to estimate the likelihood of health insurance buyers obtaining auto insurance from insurance firms. While working on the thesis, we ran into issues and challenges assessing one set of data using two algorithms. It had the most trouble deciding and picking an algorithm that could be utilised in conjunction while enhancing accuracy. It was tough to pick a specific algorithm and assess the model, as well as to figure out the alternatives for selecting and configuring various columns and running the experiment. Furthermore, throughout the process of model analysis utilising one or more algorithms, a pretty complex problem arose, causing difficulty in predicting the model. It is believed that a better model may be generated if the prediction is made using data that is more effective and has more columns than the existing one in order to produce a more accurate and fast car insurance purchase prediction algorithm model. Furthermore, the Random Forest Classifier method has the benefit of producing a quick and simple prediction model. As a result, a predictive model with improved accuracy outcomes might be developed.

VII. REFERENCE

[1] VladimirKašćelan, LjiljanaKašćelan, Milijana NovovićBurić. A nonparametric data mining approach for risk prediction in car insurance: a case study from the Montenegrin market[J]. Economic Research-Ekonomska Istraživanja,2016,29(1) :545-558.

[2] Chen M S , Hwang C P , Ho T Y , et al. Driving behaviors analysis based on feature selection and statistical approach: a preliminary study[J]. The Journal of Supercomputing, 2018.

[3] SuyeonKang, Jongwoo Song. Feature selection for continuous aggregate response and its application to auto insurance data[J]. Expert Systems With Applications,2018(93):104-117.

[4] Alshamsi, Asma S. Predicting car insurance policies using random forest[C] 2014 10th International Conference on Innovations in Information Technology (INNOVATIONS). IEEE, 2014.

[5] XiaojunMa, JinglanSha, DehuaWang, YuanboYu, QianYang, XueqiNiu. Study on A Prediction of P2P Network Loan Default Based on the Machine Learning Light GBM and XGboost Algorithms according to Different High Dimensional Data Cleaning[J]. Electronic Commerce Research and Applications, 2018(31):24-39.

[6] Yanmei Jiang, Qingkai Bu. Supermarket Commodity Sales Forecast Based on Data Mining [J]. Hans Journal of Data Mining,2018,08(02):74-78.