information

Article

Balancing Risk and Profit: Predicting the Performance of Potential New Customers in the Insurance Industry

Raquel Soriano-Gonzalez, Veronika Tsertsvadze, Celia Osorio, Noelia Fuster, Angel A. Juan and Elena Perez-Bernabeu *

Research Center on Production Management and Engineering, Universitat Politècnica de València, Ferrandiz-Carbonell, 03802 Alcoy, Spain; rsorgon@epsa.upv.es (R.S.-G.); vtserts@epsa.upv.es (V.T.); cosomuo@epsa.upv.es (C.O.); nfuscom@epsa.upv.es (N.F.); ajuanp@upv.es (A.A.J.)
* Correspondence: elenapb@upv.es

Abstract: In the financial sector, insurance companies generate large volumes of data, including policy transactions, customer interactions, and risk assessments. These historical data on established customers provide opportunities to enhance decision-making processes and offer more customized services. However, data on potential new customers are often limited, due to a lack of historical records and to legal constraints on personal data collection. Despite these limitations, accurately predicting whether a potential new customer will generate benefits (high-performance) or incur losses (low-performance) is crucial for many service companies. This study used a real-world dataset of existing car insurance customers and introduced advanced machine learning models, to predict the performance of potential new customers for whom available data are limited. We developed and evaluated approaches based on traditional binary classification models and on more advanced boosting classification models. Our computational experiments show that accurately predicting the performance of potential new customers can significantly reduce operation costs and improve the customization of services for insurance companies.

Keywords: classification of potential customers; machine learning; boosting models; insurance sector

Citation: Soriano-Gonzalez, R.; Tsertsvadze, V.; Osorio, C.; Fuster, N.; Juan, A.A.; Perez-Bernabeu, E. Balancing Risk and Profit: Predicting the Performance of Potential New Customers in the Insurance Industry. *Information* **2024**, *15*, 546. https://doi.org/10.3390/info15090546

Academic Editors: Polona Tominc and Maja Rožman

Received: 31 July 2024
Revised: 3 September 2024
Accepted: 4 September 2024
Published: 6 September 2024

1. Introduction

Machine learning (ML) has rapidly expanded in recent years, as indicated by numerous studies [1–3]. This growth is primarily due to the increasing availability of data and the ability of ML to process, prepare, and analyze extensive datasets. The contributions of new statistical computing methods and big data tools for managing and understanding large datasets is remarkable in this context [4]. Additionally, several researchers have highlighted the potential of cloud computing technologies to process vast amounts of data and the challenges associated with managing this volume of data [5–7]. Similarly, other studies have emphasized the significance of data exploration and visualization systems for interpreting and extracting information from large datasets [8]. These studies show the need for advanced technologies and methods to handle, prepare, and analyze large datasets efficiently [9]. The application of these technologies spans various sectors, such as manufacturing, finance, commerce, and insurance, allowing more evidence-based decision-making processes [10]. Focusing on the car insurance industry, there has been a notable shift towards digital applications and increased use of ML techniques, due to the rising number and severity of auto insurance claims. This trend calls for new methods to manage these claims efficiently with ML predictive models [11,12]. As customer data volumes grow, there are significant opportunities to improve methodologies, such as policy enrollment, claims settlement processes, and understanding customer behavior [13]. Despite ML's advantages to this industry, its full potential remains unexploited. One common activity for many firms involves segmentation of already established customers, using their predictive

behavior and risk assessment [14–16]. While segmentation of existing customers is well-documented and advanced [17], predicting the performance of potential new customers, for whom historical data are unavailable, remains a challenging task. Therefore, a gap exists in developing ML models that can accurately classify potential new customers based on their performance. The challenge with new customers is that only minimal information or data are usually available, making classification difficult. The task then becomes finding ways to classify these potential new customers with limited information. This indicates that existing tools and methods may fall short in this regard [18]. Therefore, there is significant room for improvement in the classification of potential new customers within the insurance industry. An additional critical but often overlooked aspect in the literature is the cost of misclassification. Misclassifying customers can lead to significant financial repercussions for insurance companies. For example, misclassifying a true low-performance new customer as a high-performance one (a false negative) can result in underpriced premiums that do not cover future claims. Conversely, misclassifying a true high-performance new customer as a low-performance one (a false positive) can lead to overpriced premiums, potentially driving valuable customers to competitors [19].

In this context, the main contribution of this paper is to propose and compare a series of traditional and advanced classification models that can be employed over a real-life dataset, to predict the performance of potential new customers in a car insurance company. This approach represents a novel contribution to the field, aiming to enhance the precision and effectiveness of new customer evaluation. To accomplish this, the paper is structured as follows: Section 1 introduces the topic by outlining the advantages of applying ML to businesses, emphasizing its impacts within the insurance industry. This section also identifies existing gaps and the unexploited potential of ML in the classification of potential new customers with limited data availability, which is the main focus of this research. Section 2 reviews the relevant literature, providing a foundation for the methodology considered. Section 3 details the data preprocessing steps, describes the target variable, and explains the framework for developing the proposed models. Using cross-validation techniques, Section 4 offers a comparison of a pool of classification models, which ranges from traditional ones to more recent ones. Section 5 performs a similar analysis for a selected subset of advanced models, but this time employing an additional validation dataset and early stopping to avoid overfitting. Section 6 presents the outcomes of the selected model, demonstrating its effectiveness and limitations for classifying potential new customers. Finally, Section 7 summarizes the findings, highlights the impact of using ML models for predicting the performance of potential new customers in the insurance sector, and suggests potential future research directions.

2. Literature Review

The integration of ML in the financial sector, particularly in insurance, has received significant attention, due to its potential to improve risk management and operational efficiency. The ability of ML to handle large datasets and perform complex analyses is well-documented, with numerous studies highlighting its benefits in processing large datasets, facilitating exploratory analysis, classification, and predictive analytics [20–22].

Research has demonstrated the application of ML algorithms, such as support vector machines, random forest, and naive Bayes, to predicting the risk levels of prospective customers, thus reducing manual effort in underwriting and optimizing resource allocation [23]. Similarly, studies have explored the application of ML to managing auto insurance claims, showing improved predictive models through the use of big data [12]. Customer segmentation involves using clustering techniques, such as k-means, hierarchical clustering, and DBSCAN, to group customers based on shared characteristics, facilitating targeted marketing and customer service strategies [24]. However, a critical gap in the literature exists in the application of these models to scenarios where data availability is limited or where the cost of acquiring data is prohibitive. Although customer segmentation is well-established, classifying potential new customers, especially those with minimal

available data, remains underexplored [25]. The challenge of predicting the performance potential of customers with limited information is significant, as existing models often do not fully exploit the patterns and trends that can be achieved for existing customers with historical records. This issue becomes more pronounced in cases where the economic and financial implications of misclassification are substantial. Yet, few studies have addressed how to mitigate these risks cost-effectively [11].

In addition to clustering techniques, ensemble methods have shown effectiveness in improving predictive performance. Techniques like bagging, boosting, and stacking combine the strengths of multiple models to enhance classification accuracy and robustness [26]. These ensemble methods have been applied in various insurance-related tasks, from fraud detection to claim prediction, demonstrating their utility in different contexts [27]. Moreover, the advent of deep learning has further expanded ML applications in the insurance industry. Convolutional neural networks and recurrent neural networks, including advanced variants like long short-term memory networks, have been utilized for tasks such as image and sequential data analysis, respectively [28]. These techniques enable insurers to analyze unstructured data, such as images of damaged vehicles, for claims processing, with notable accuracy and efficiency [29,30]. Integrating natural language processing (NLP) techniques with ML models has also shown promising results in processing textual data from customer interactions, policy documents, and claim descriptions [31–33].

Despite these advances, challenges remain in the application of ML in insurance. One significant issue is the interpretability of ML models, particularly complex ones, such as deep learning models. Insurers must ensure that their ML models are accurate and interpretable, to gain regulatory approval and maintain customer trust [34,35]. Techniques such as SHAP (Shapley additive explanations) and LIME (local interpretable model-agnostic explanations) have been developed to address this challenge by providing interpretable explanations of model predictions [36,37]. Another challenge is data privacy and security. Insurance companies handle sensitive customer data, and implementing ML models must comply with data protection regulations, such as the general data protection regulation in the European Union [38,39].

3. Overview of Methodology and First Steps

The proposed model is designed to guide decision making in an insurance company, aiming to predict the performance of potential new customers effectively. This approach helps the company better identify potential new customers who may generate future benefits (high-performance) or losses (low-performance). Hence, when potential new customers apply for or express interest in insurance policies, a high-quality classification model can predict their performance and help offer customized policies based on these predictions. The methodological process for developing this model is visually summarized in Figure 1. This figure represents the methodology workflow, from the initial selection of variables to the validation of classification models. Following this approach, this section describes the study's methodology in three main steps: (i) data preparation; (ii) variable selection; and (iii) models development.

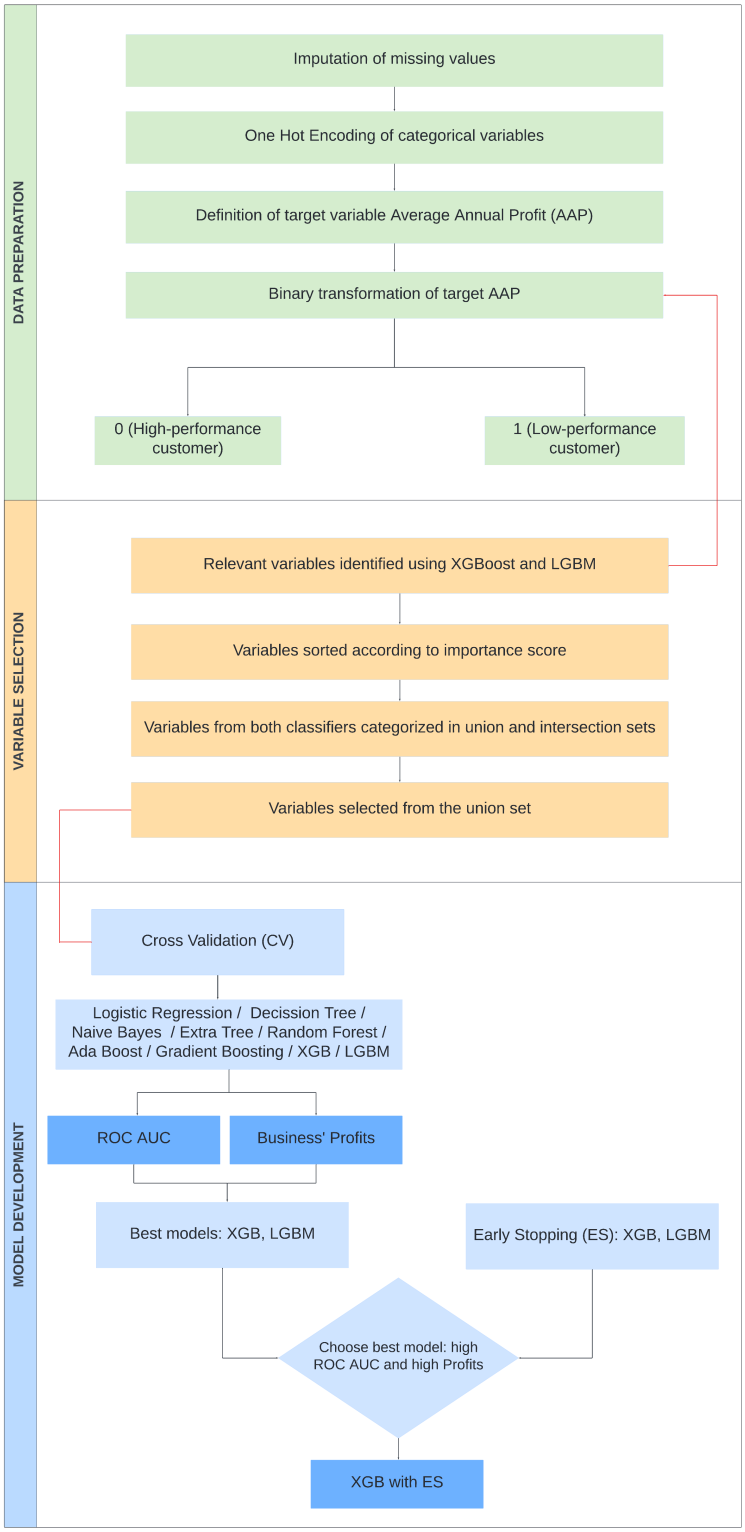

Figure 1. Methodology workflow of the proposed model.

3.1. Data Preparation

The dataset used in this study consists of 116,934 records of established insurance customer data from 2016 to 2023, including personal characteristics, vehicle attributes, and accident records from the past five years. Before proceeding with the data analysis, several preprocessing steps were necessary, to prepare the dataset for effective use. These preprocessing steps are summarized in Table 1.

Table 1. Steps performed during data preprocessing.

Process	Description	Method
Handling missing values	Imputation of missing data, to ensure dataset completeness and integrity.	The 'soft impute' method was employed; it ensured that the imputed values maintained the inherent patterns and relationships within the dataset, minimizing distortion of the original data distribution [40,41].
Outlier management	Identification and handling of anomalous records, to refine data quality.	Removal of fleet data and review of anomalous values.
Date variable transformation	Transformation of date variables to year format, to streamline analysis and simplify temporal data handling.	Conversion to year format.
Categorical variable encoding	Transformation of categorical variables into a numerical format, to facilitate model interpretation.	By using the 'one-hot encoding' method [42], each vehicle type category was substituted by the corresponding dummy variables without imposing hierarchy or order on them.

The next stage was to create new variables. This included defining the target variable 'average annual profit' (AAP), using a formula proposed by the firm, which incorporated the values of premiums, claims, and commissions and the insured's exposure. The formula is expressed in Equation (1):

$$\text{AAP} = \frac{p - c - a - kp}{e} \tag{1}$$

where the variables are as follows:

- p refers to the premiums or total amount paid by the established customer for the insurance policies.
- c represents the claims recorded on the policy from the policy start date.
- a denotes the amount paid to agents as a percentage of the premiums.
- The constant k represents the portion of the premiums allocated by the insurance company to cover administrative expenses (in our case, $k = 0.15$).
- e represents the percentage of exposure, i.e., the percentage of days in a year that the policy has been active.

Columns (other than AAP) that could not be obtained for potential new customers were removed from the original dataset. Likewise, records without SINCO values were removed as well, where SINCO refers to the information system of the insurance compensation consortium in Spain. After these modifications, the resulting dataset included 51,618 observations and 196 variables. Approximately 19% of the observations corresponded to true low-performance customers (AAP ≤ 0, true target $= 1$), with the remaining 81% of the observations corresponding to true high-performance customers (AAP > 0, true target $= 0$). Table 2 presents some of these variables, where INE refers to the Spanish National Institute of Statistics and DGT refers to the Spanish General Directorate of Traffic.

Table 2. Some of the main variables included in the final dataset.

Variable	Description	Data Type	Transformation
Age	The customer's age in years at the time the policy is taken out.	Discrete	From date format to discrete number.
Risk	Level of associated risk, generated from the equalities and inequalities between the driver, the vehicle owner, and the policy payer.	Categorical	Elimination of driver, owner, and co-driver identification variables.
Horse power	The power of the vehicle's engine.	Discrete	N/A
Historical family premiums	Accumulated net premiums of policies contracted by the customer's family members in the company.	Continuous	N/A
License age	The number of years the customer has held a driver's license.	Discrete	Calculated from the variables card issue date and policy application date.
Hiring channel score	Score assigned to the salesperson in charge of the policy contracting process.	Categorical	N/A
INE income	Represents the average annual net income of the region where the customer resides, according to INE.	Continuous	The missing values have been imputed using the soft impute technique.
DGT accident	The average annual number of accidents with fatalities recorded by the DGT, value computed for the average 2016–2022.	Continuous	The missing values have been imputed using the soft impute technique.
Exposure	Duration of the customer's insurance in the last 5 years, extracted from the SINCO database.	Discrete	Calculated from the dates of the policies registered in SINCO.
Frequency of material damage	Frequency of accidents with material damage obtained after consulting SINCO.	Continuous	Calculated from the dates of occurrence of this type of accident.
Frequency of personal damage	Frequency of accidents with personal damage obtained after consulting SINCO.	Continuous	Calculated from the dates of occurrence of this type of accident.
Claim outcome score	Score assessing the claims obtained from SINCO.	Discrete	N/A

Figure 2 shows two boxplots of AAP values in euros: one with outliers and one without. In the boxplot with outliers, the boxes are concentrated around 0, and the presence of extreme outliers makes the actual boxes difficult to see. Likewise, some additional statistics on the AAP variable are provided in Table 3. While the vast majority of customers had a positive AAP value (high-performance), a few customers generated significant losses for the company. Predicting the performance of these customers in advance would be highly beneficial for the firm, as it could establish customized policies for them and reduce the operational costs associated with future large claims.

Table 3. Summary of statistics for the AAP variable on the entire dataset.

Statistic	Value
Count	51,618 customers
Mean	18 euros
StDev	2277 euros
Min	−185,990 euros
Q1	73 euros
Median	178 euros
Q3	262 euros
Max	21,747 euros

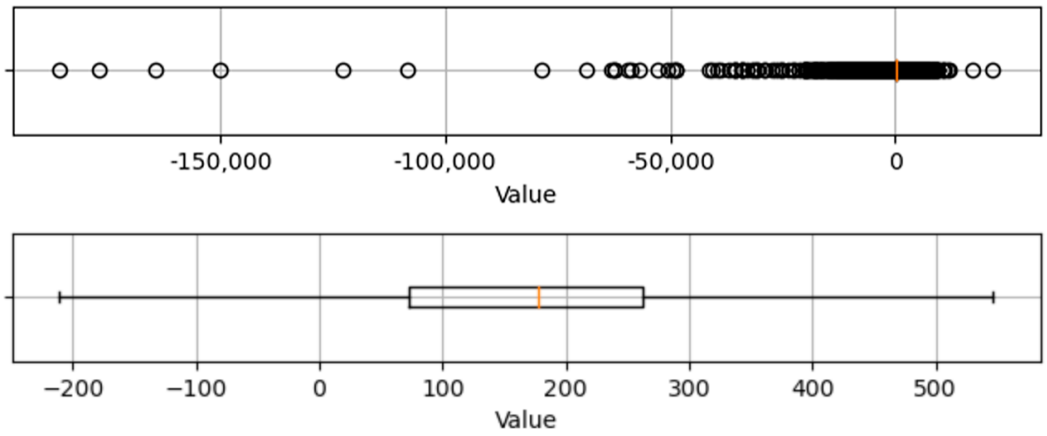

Figure 2. Boxplots of AAP values (in euros) with and without outliers.

In order to create a binary target variable (high-performance vs. low-performance customer), the APP values were then transformed from real values to binary ones based on a predefined threshold: instances where AAP was positive were classified as 0, indicating a high-performance customer, and those with a negative or zero APP value were classified as 1, indicating a low-performance customer.

3.2. Selection of Relevant Features

After describing the dataset preparation process, the selection of relevant features was conducted, using the following two classifiers: extreme gradient boosting (XGBoost or XGB) [43] and light gradient-boosting machine (LightGBM or LGBM) [44]. On the one hand, XGB uses a gradient boosting framework optimized for both efficiency and performance, particularly suitable for large-scale predictive modeling [45]. The algorithm assigns a normalized importance score to each feature, quantifying its relative contribution to model accuracy. This score ranges from 0 to 1, with each fraction representing the proportion of the model's predictive power attributed to that particular variable [43]. On the other hand, LGBM provides a complementary approach with its distributed boosting algorithm. It measures the importance of each feature based on how frequently the variable is used to make splits in the decision trees throughout the model [46]. In this case, a high frequency indicates an important impact on the outcome [44]. Given the advantages offered by these algorithms, they were used to evaluate and sort variables based on their importance scores. The selected variables and importance scores are displayed for each classifier. Figure 3 shows that only a subset of seven features, from the original 196 variables, were considered as relevant (importance score ≥ 0.01) according to the XGB classifier.

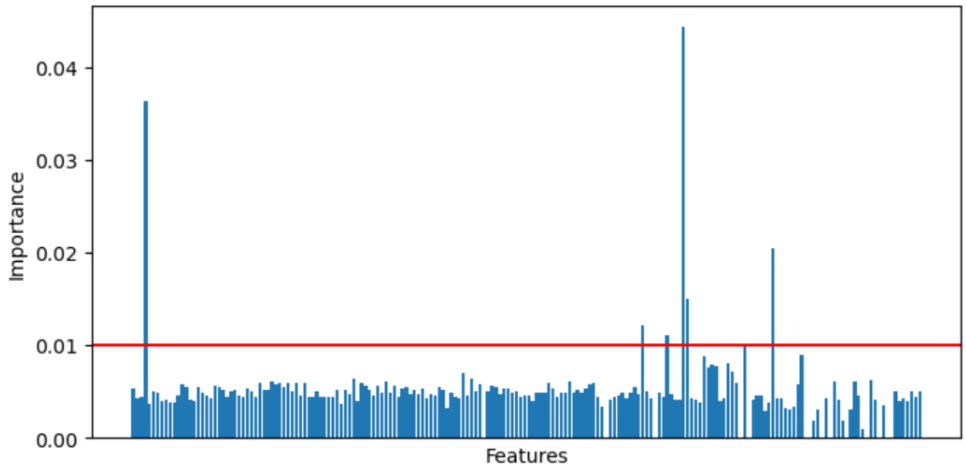

Figure 3. Important features according to the XGB classifier.

The final step involved identifying both common and unique relevant features across the XGB and LGBM classifiers. This was accomplished by analyzing the union and intersection of the relevant variables obtained from both models. The union set compiled all the unique features considered relevant by any of the classifiers, while the intersection set gathered those features consistently identified across both classifiers. Figure 4 illustrates the results obtained from evaluating the importance of the variables, using both XGB and LGBM classifiers. It also shows the intersection of the variable importance scores derived from these models, providing insights into which variables were significant across both classifiers. Relevant features included several variables of customer demographics, vehicle information, insurance history, and claims frequency from SINCO. After considering the variables obtained from the union and intersection sets, the decision was made to select the variables in the union set rather than those in the intersection. By adopting the union criterion for variable selection, the focus was on using the most complete set of variables considered relevant by either classifier.

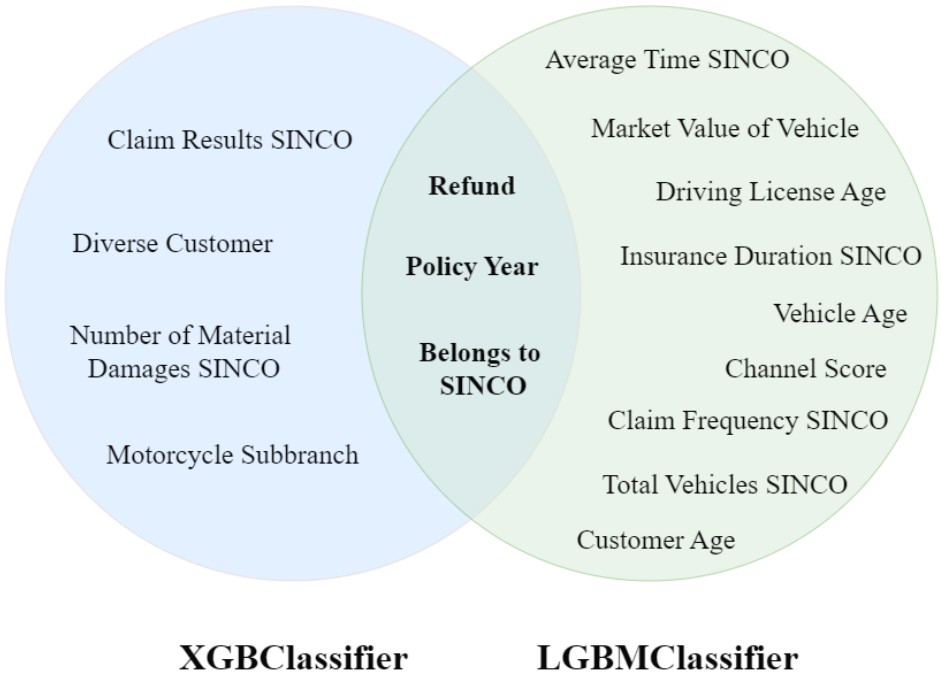

Figure 4. Relevant variables identified by XGB and LGBM classifiers.

4. Evaluation of a Pool of Models Using Cross-Validation

After selecting the relevant variables, several traditional and advanced classification models were considered, to improve predictive accuracy. First, the dataset was divided into a training set (70%) and a test set (30%), to perform a cross-validation analysis. A diverse set of classification algorithms was selected. These models were chosen to address different data dynamics hypotheses and to explore several statistical and computational approaches. This initial pool of models included traditional algorithms, such as logistic regression [47], decision trees [48], naive Bayes [49], extra trees, random forest, Ada boost, and gradient boosting. More recent ensemble techniques, such as XGB and LGBM, were also included. XGB [43] is a gradient-boosting framework that optimizes custom loss functions and is highly effective for large datasets. LGBM [44] is known for its speed and high performance, handling large data volumes efficiently. Using the training data, the performance and ability of the individual models to generalize to unseen data were quantified, using the receiver operating characteristic area under the curve (ROC AUC) score. This metric, which scales from 0 to 1, evaluates the ability of a model to distinguish between classes. A higher ROC AUC score indicates better discriminatory power [50]. The ROC AUC score was specifically chosen in this study because it balances identifying potential customers (sensitivity) with

correctly excluding non-potential ones (specificity). This balance helps prevent the financial negative effects for the insurance company of customer misclassification [51]. Table 4 shows a comparison of the aforementioned classification models using cross-validation, where the additional columns indicate the following: (i) the number and percentage of potential new customers from the test dataset that would be classified as low-performance ones; (ii) the estimated benefit (in euros) associated with the test dataset; and (iii) the execution time, in seconds, employed to execute the model on a standard laptop computer running Linux Mint 21.1 with an Intel *i*5-7200*U*(4)@3.1 GHz CPU and 16 GB RAM.

Table 4. Comparison of different classification models, using cross-validation.

Model	ROC AUC (Train)	Low Perform. (Test)	Est. Benefit (Test)	Time (s)
naive Bayes	0.624	1462 (9%)	1,092,228 €	9
decision trees	0.633	1634 (11%)	1,228,023 €	9
logistic regression	0.653	194 (1%)	513,173 €	11
multi-layer perceptron	0.660	48 (0%)	445,084 €	18
extra trees	0.669	1583 (10%)	1,249,698 €	20
random forest	0.687	1571 (10%)	1,270,656 €	19
adaboost	0.697	264 (2%)	611,936 €	12
XGB	0.709	1099 (7%)	1,269,588 €	10
gradient boosting	0.713	973 (6%)	1,183,631 €	15
LGBM	0.717	1000 (6%)	1,226,696 €	10

Note that, in terms of the ROC AUC score for the train dataset, the best models were LGBM, gradient boosting, and XGB, all of them with a score around 0.71. In particular, both the LGBM and the XGB models seeme to exhibit excellent performance regarding several dimensions that were requested by the firm: (i) a low percentage of customers who were classified as low-performance (6% and 7%, respectively); (ii) a relatively high estimated benefit when the model was applied on the test dataset, which was quite high in both cases (1,226,696 euros and 1,269,588 euros, respectively), especially when compared with the estimated benefit of just 406,006 euros the firm would obtain without using these models; and (iii) the low computational times (10 s) required to execute the entire model, even when using a standard computer. Even when a greedy search of parameters was run on these two models, increases in the selected score were small and did not justify the extra complexity and computational time required.

5. Evaluation of Selected Models, Using Early Stopping

This section presents a complementary analysis using a different approach, which was based on the utilization of a validation set and the early stopping technique applied to the XGB and LGBM classifiers. These two classifiers were selected due to their excellent performance in the previous set of computational experiments. Thus, now the dataset of 51,618 observations was split first into two subsets: a training–validation one (70%) and a test one (30%). Next, the training and validation set was split again into a training set (90%) and a validation set (10%), to be used during the early stopping procedure. Table 5 shows a comparison of the aforementioned classification models, using a validation set and early stopping. The first columns show the ROC AUC scores for the training and validation sets, respectively. As before, the additional columns indicate the following: (i) the number and percentage of potential new customers from the test dataset that would be classified as low-performance; (ii) the estimated benefit (in euros) associated with the test dataset; and (iii) the execution time, in seconds, employed to execute the model in the previous standard laptop computer.

Table 5. Comparison of advanced classification models using early stopping.

Model	ROC AUC (Train)	ROC AUC (Val)	Low Perform. (Test)	Est. Benefit (Test)	Time (s)
LGBM	0.738	0.723	423 (3%)	763,145 €	8
XGB	0.748	0.742	1003 (6%)	1,232,663 €	10

From the results provided in Tables 4 and 5, it seems clear that the proposed XGB models demonstrated solid performance in both the cross-validation and early stopping scenarios. Specifically, in the early stopping scenario, the ROC AUC score reached 0.74 for the validation set, with an estimated benefit of 1,232,663 euros when applied to the test set. These results are notable, especially considering that (i) they pertain to potential new customers for whom limited data were available, and (ii) the estimated benefit to the company for the test set would be around 406,006 euros if these models were not used.

6. Further Analysis on the Test Dataset

Based on the previous experiments and the results obtained for the training and validation datasets, the XGB model with early stopping was selected and applied to the test dataset, which consisted of 15,486 customers (30% of the initial 51,618 observations that represented potential new customers). From these test observations, the model classified 1003 customers (6% of the test observations) as low-performance, while the remaining 14,483 customers (94% of the test observations) were classified as high-performance. Table 6 displays summary statistics for the set of customers that were predicted by the model as high-performance, while Figure 5 shows boxplots for the APP values (in euros) associated with customers classified as high-performance (0 or HP) and low-performance (1 or LP). As shown in the boxplots, some customers classified as high-performance were actually low-performance (false negatives), as indicated by their negative AAP values. Conversely, a few customers classified as low-performance were, in reality, high-performance with positive AAP values (false positives). These misclassification mistakes were expected, due to the challenging task of predicting the performance of potential new customers. Despite this, the average AAP value for the customers classified by the model as high-performance was 85 euros (Table 6), which represented a noticeable increase with respect to the average APP value of 18 euros obtained before applying the classification model to filter out potential low-performance customers (Table 3).

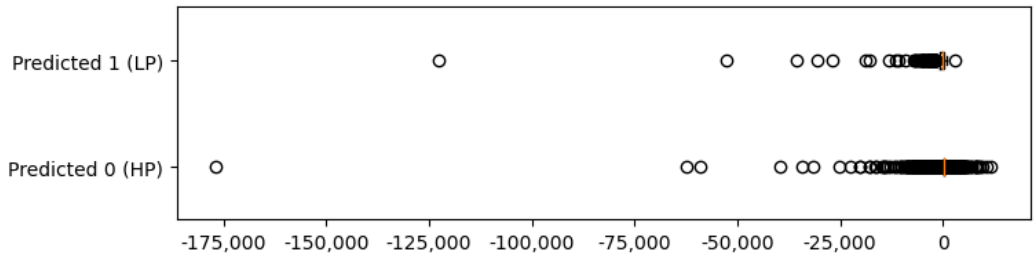

Figure 5. Boxplots of AAP values (in euros) for customers classified as high-performance (HP) and low-performance (LP).

Table 6. Summary of statistics for the AAP variable on the predicted high-performance customers.

Statistic	Value
Count	14,483 customers
Mean	85 euros
StDev	1964 euros
Min	−176,973 euros
Q1	92 euros

Table 6. *Cont.*

Statistic	Value
Median	183 euros
Q3	267 euros
Max	11,774 euros

Figure 6 displays the confusion matrix of the model for the test dataset, providing additional details on the number of correctly classified customers as well as false positives and false negatives. Note that, out of the 14,483 customers in the test dataset, 2296 were mistakenly classified as high-performance, and just 326 were misclassified as low-performance.

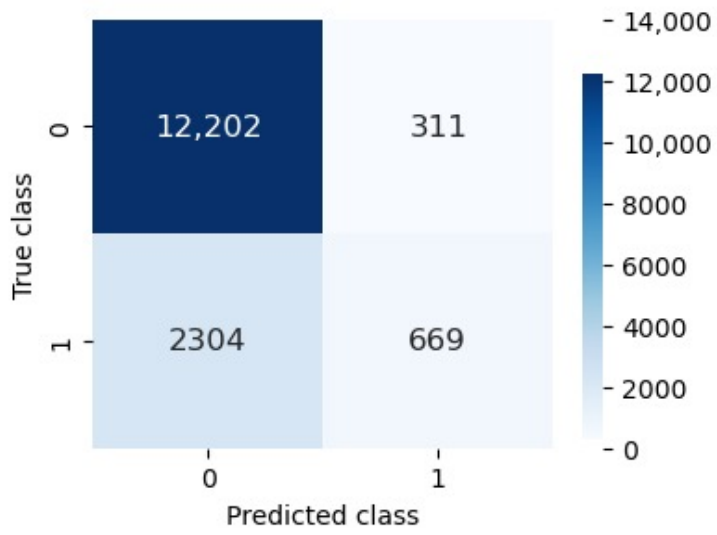

Figure 6. Confusion matrix of the selected model for the test dataset.

Table 7 provides the classification report for the test dataset. Note that while the precision, recall, and F1 scores were high for class 0 (high-performance), they were relatively low for class 1 (low-performance). This discrepancy was due to the challenge of predicting future low-performance customers among potential new customers who had no historical records with the company.

Table 7. Classification report for the test dataset.

Class	Precision	Recall	F1-Score	Support
0	0.84	0.97	0.90	12,513
1	0.67	0.23	0.34	2973
Accuracy		0.83		15,486
Macro Avg	0.76	0.60	0.62	15,486
Weighted Avg	0.81	0.83	0.79	15,486

Figure 7 displays the distribution of the AAP values that would be lost due to the misclassification of 326 high-performance customers (false positives). The average AAP value for these misclassified customers would be around 228 euros, with a standard deviation of 223 euros, a minimum value of 2 euros, a first quartile of 111 euros, a median of 196 euros, a third quartile of 271 euros, and a maximum value of 2947 euros.

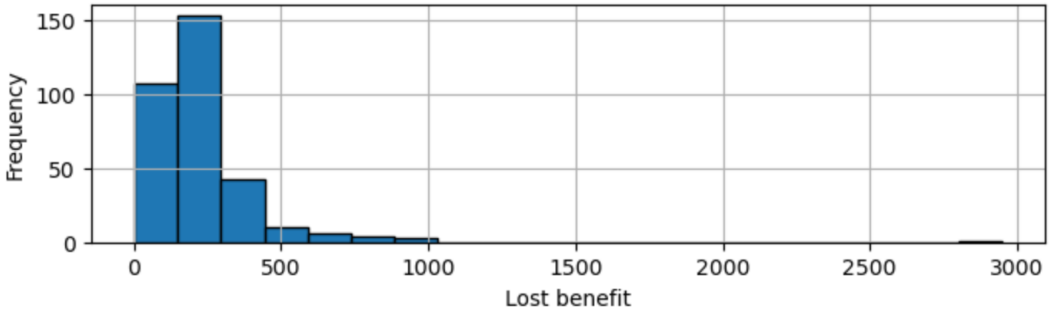

Figure 7. Lost benefit for erroneously misclassifying high-performance customers.

Finally, the ROC curve in Figure 8 shows the trade-off between the true and false positive rates at various threshold settings. The ROC AUC score of 0.72 indicates that the model has a reasonable ability to distinguish between potential high-performance customers and other customers. Overall, by utilizing the proposed model to identify potential low-performance customers in the test dataset, the firm could reduce its operational costs and, consequently, increase its estimated benefit to 1,232,663 euros. This represents a significant increase compared to the current benefit of 406,006 euros. Thus, by using the proposed model to identify potential high-performance and low-performance new customers, the company could reduce the risk of incurring extremely costly future claims, and it could maintain a sustainable business model.

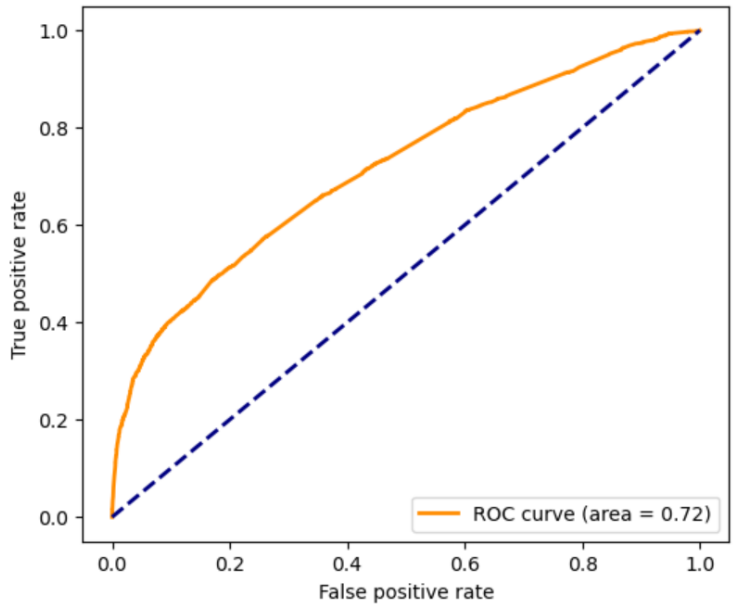

Figure 8. ROC curve for XGB model with ES on the test dataset.

7. Conclusions

After identifying a gap in the literature regarding the prediction of performance for potential new customers with limited available data, this study developed and tested various classification models, using real-life data from an insurance company. The results show that advanced models based on XGB or LGBM classifiers generally outperform traditional models, such as logistic regression, decision trees, naive Bayes, and random forest. Specifically, our computational experiments demonstrate that the XGB model, utilizing validation data and early stopping to reduce the risk of overfitting, performs reasonably well, even for potential new customers with limited data.

Although metrics such as ROC AUC and F1 scores did not reach extremely high levels—mainly due to the lack of historical data for potential new customers—the true value of the model lies in its ability to accurately classify low-performance customers who

could represent significant financial losses for the insurance company. The application of the model increased the company's estimated benefits for the test dataset of 14,483 customers, from 406,006 euros to 1,232,663 euros. Thus, the model improved the classification process for potential new customers, enabling the application of customized policies. The analysis highlights the need for improved prediction in insurance companies, particularly in identifying low-performance customers, as errors in this classification can significantly increase operational costs.

Future research includes exploring more advanced deep learning models, which may yield better ROC AUC and F1 scores. Additionally, we are working with the firm to identify new variables for potential new customers that align with the current regulatory framework. Finally, we plan to extend these models to other types of policies within the company.

Author Contributions: Conceptualization, E.P.-B. and A.A.J.; methodology, A.A.J.; software, A.A.J.; validation, R.S.-G. and V.T.; formal analysis, R.S.-G.; data curation, N.F.; writing—original draft preparation, R.S.-G., C.O., V.T., N.F. and A.A.J.; writing—review and editing, A.A.J. and E.P.-B.; visualization, V.T. and R.S.-G.; supervision, E.P.-B. All authors have read and agreed to the published version of the manuscript.

Funding: This work was partially funded by the Investigo Program of the Generalitat Valenciana (INVEST/2023/304).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets presented in this article are unavailable because they are a company's private data.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ML	Machine Learning
ROC AUC	Receiver Operating Characteristic Area Under the Curve
AAP	Average Annual Profit

References

1. Krenn, M.; Buffoni, L.; Coutinho, B.; Eppel, S.; Foster, J.G.; Gritsevskiy, A.; Lee, H.; Lu, Y.; Moutinho, J.P.; Sanjabi, N.; et al. Forecasting the future of artificial intelligence with machine learning-based link prediction in an exponentially growing knowledge network. *Nat. Mach. Intell.* **2023**, *5*, 1326–1335. [CrossRef]
2. Dinov, I.D. Volume and value of big healthcare data. *J. Med. Stat. Inform.* **2016**, *4*, 3. [CrossRef] [PubMed]
3. Jordan, M.I.; Mitchell, T.M. Machine learning: Trends, perspectives, and prospects. *Science* **2015**, *349*, 255–260. [CrossRef] [PubMed]
4. Kuznetsov, V. Gaining insight from large data volumes with ease. In *EPJ Web of Conferences*; EDP Sciences: Sofia, Bulgaria, 2019; Volume 214, p. 04027.
5. Rani, S.; Bhambri, P.; Kataria, A. Integration of IoT, Big Data, and Cloud Computing Technologies: Trend of the Era. In *Big Data, Cloud Computing and IoT*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2023; pp. 1–21.
6. Ionescu, S.A.; Diaconita, V. Transforming financial decision-making: The interplay of AI, cloud computing and advanced data management technologies. *Int. J. Comput. Commun. Control* **2023**, *18*, 5735. [CrossRef]
7. Siddiqa, A.; Hashem, I.A.T.; Yaqoob, I.; Marjani, M.; Shamshirband, S.; Gani, A.; Nasaruddin, F. A survey of big data management: Taxonomy and state-of-the-art. *J. Netw. Comput. Appl.* **2016**, *71*, 151–166. [CrossRef]
8. Raghav, R.S.; Pothula, S.; Vengattaraman, T.; Ponnurangam, D. A survey of data visualization tools for analyzing large volume of data in big data platform. In Proceedings of the 2016 International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 21–22 October 2016; pp. 1–6.
9. Jones, K.I.; Sah, S. The Implementation of Machine Learning In The Insurance Industry With Big Data Analytics. *Int. J. Data Inform. Intell. Comput.* **2023**, *2*, 21–38.
10. Jamal, S.; Goyal, S.; Grover, A.; Shanker, A. Machine Learning: What, Why, and How? In *Bioinformatics: Sequences, Structures, Phylogeny*; Springer: Singapore, 2018; pp. 359–374.

11. Tian, X.; Todorovic, J.; Todorovic, Z. A Machine-Learning-Based Business Analytical System for Insurance Customer Relationship Management and Cross-Selling. *J. Appl. Bus. Econ.* **2023**, *25*, 273. [CrossRef]
12. Hanafy, M.; Ming, R. Machine learning approaches for auto insurance big data. *Risks* **2021**, *9*, 42. [CrossRef]
13. Rawat, S.; Rawat, A.; Kumar, D.; Sabitha, A.S. Application of machine learning and data visualization techniques for decision support in the insurance sector. *Int. J. Inf. Manag. Data Insights* **2021**, *1*, 100012. [CrossRef]
14. Mahbobi, M.; Kimiagari, S.; Vasudevan, M. Credit risk classification: An integrated predictive accuracy algorithm using artificial and deep neural networks. *Ann. Oper. Res.* **2023**, *330*, 609–637. [CrossRef]
15. Hosein, P. A data science approach to risk assessment for automobile insurance policies. *Int. J. Data Sci. Anal.* **2024**, *17*, 127–138. [CrossRef]
16. Jeong, H.; An, J.; Jeong, J. Are you a good client? Client classification in federated learning. In Proceedings of the 2021 International Conference on Information and Communication Technology Convergence (ICTC), Jeju Island, Republic of Korea, 20–22 October 2021; pp. 1691–1696.
17. Eluwole, O.T.; Akande, S. Artificial Intelligence in Finance: Possibilities and Threats. In Proceedings of the 2022 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), Virtual, 28–30 July 2022; pp. 268–273.
18. Luciano, E.; Cattaneo, M.; Kenett, R. Adversarial AI in Insurance: Pervasiveness and Resilience. *arXiv* **2023**, arXiv:2301.07520.
19. Finger, D.; Albrecher, H.; Wilhelmy, L. On the cost of risk misspecification in insurance pricing. *Jpn. J. Stat. Data Sci.* **2024**, 1–43. [CrossRef]
20. Leo, M.; Sharma, S.; Maddulety, K. Machine learning in banking risk management: A literature review. *Risks* **2019**, *7*, 29. [CrossRef]
21. Fitriani, M.A.; Febrianto, D.C. Data mining for potential customer segmentation in the marketing bank dataset. *JUITA J. Inform.* **2021**, *9*, 25–32. [CrossRef]
22. Simester, D.; Timoshenko, A.; Zoumpoulis, S.I. Targeting prospective customers: Robustness of machine-learning methods to typical data challenges. *Manag. Sci.* **2020**, *66*, 2495–2522. [CrossRef]
23. Hutagaol, B.J.; Mauritsius, T. Risk level prediction of life insurance applicant using machine learning. *Int. J. Adv. Trends Comput. Sci. Eng.* **2020**, *9*, 2213–2220.
24. Sadreddini, Z.; Donmez, I.; Yanikomeroglu, H. Cancel-for-Any-Reason Insurance Recommendation Using Customer Transaction-Based Clustering. *IEEE Access* **2021**, *9*, 39363–39374. [CrossRef]
25. Sari, P.K.; Purwadinata, A. Analysis characteristics of car sales in E-commerce data using clustering model. *J. Data Sci. Appl.* **2019**, *2*, 19–28. [CrossRef]
26. Dietterich, T.G. Ensemble methods in machine learning. In Proceedings of the International Workshop on Multiple Classifier Systems, Cagliari, Italy, 21–23 June 2000; Springer: Berlin/Heidelberg, Germany, 2000; pp. 1–15.
27. Zhou, Z.H. *Ensemble Methods: Foundations and Algorithms*; CRC Press: Boca Raton, FL, USA, 2012.
28. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
29. Elbhrawy, A.S.; Belal, M.A.; Hassanein, M.S. CES: Cost Estimation System for Enhancing the Processing of Car Insurance Claims. *J. Comput. Commun.* **2024**, *3*, 55–69. [CrossRef]
30. De Meulemeester, H.; De Moor, B. Unsupervised embeddings for categorical variables. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; pp. 1–8.
31. Kolambe, S.; Kaur, P. Survey on Insurance Claim analysis using Natural Language Processing and Machine Learning. *Int. J. Recent Innov. Trends Comput. Commun.* **2023**, *11*, 30–38. [CrossRef]
32. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75. [CrossRef]
33. Cambria, E.; White, B. Jumping NLP curves: A review of natural language processing research. *IEEE Comput. Intell. Mag.* **2014**, *9*, 48–57. [CrossRef]
34. Orji, U.; Ukwandu, E. Machine learning for an explainable cost prediction of medical insurance. *Mach. Learn. Appl.* **2024**, *15*, 100516. [CrossRef]
35. Doshi-Velez, F.; Kim, B. Towards a rigorous science of interpretable machine learning. *arXiv* **2017**, arXiv:1702.08608.
36. Le, T.T.H.; Prihatno, A.T.; Oktian, Y.E.; Kang, H.; Kim, H. Exploring local explanation of practical industrial AI applications: A systematic literature review. *Appl. Sci.* **2023**, *13*, 5809. [CrossRef]
37. Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 4765–4774.
38. Sharma, A. Demystifying Privacy-preserving AI: Strategies for Responsible Data Handling. *MZ J. Artif. Intell.* **2024**, *1*, 1–8.
39. Voigt, P.; Von dem Bussche, A. The eu general data protection regulation (gdpr). In *A Practical Guide*, 1st ed.; Springer International Publishing: Cham, Switzerland, 2017; Volume 10.
40. Hastie, T.; Mazumder, R.; Lee, J.D.; Zadeh, R. Matrix completion and low-rank SVD via fast alternating least squares. *J. Mach. Learn. Res.* **2015**, *16*, 3367–3402.
41. Rafsunjani, S.; Safa, R.S.; Al Imran, A.; Rahim, M.S.; Nandi, D. An empirical comparison of missing value imputation techniques on APS failure prediction. *Int. J. Inf. Technol. Comput. Sci.* **2019**, *2*, 21–29. [CrossRef]
42. Hancock, J.; Khoshgoftaar, T.M. Leveraging lightgbm for categorical big data. In Proceedings of the 2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService), Virtual, 23–26 August 2021; pp. 149–154.

43. Li, S.; Zhang, X. Research on orthopedic auxiliary classification and prediction model based on XGBoost algorithm. *Neural Comput. Appl.* **2020**, *32*, 1971–1979. [CrossRef]
44. Alzamzami, F.; Hoda, M.; El Saddik, A. Light gradient boosting machine for general sentiment classification on short texts: A comparative evaluation. *IEEE Access* **2020**, *8*, 101840–101858. [CrossRef]
45. Abdurrahman, G.; Sintawati, M. Implementation of Xgboost for Classification of Parkinson's Disease. *J. Phys. Conf. Ser.* **2020**, *1538*, 012024. [CrossRef]
46. Sari, L.; Romadloni, A.; Lityaningrum, R.; Hastuti, H.D. Implementation of LightGBM and Random Forest in Potential Customer Classification. *TIERS Inf. Technol. J.* **2023**, *4*, 43–55. [CrossRef]
47. Dreiseitl, S.; Ohno-Machado, L. Logistic regression and artificial neural network classification models: A methodology review. *J. Biomed. Inform.* **2002**, *35*, 352–359. [CrossRef]
48. Charbuty, B.; Abdulazeez, A. Classification based on decision tree algorithm for machine learning. *J. Appl. Sci. Technol. Trends* **2021**, *2*, 20–28. [CrossRef]
49. Gladence, L.M.; Karthi, M.; Anu, V.M. A statistical comparison of logistic regression and different Bayes classification methods for machine learning. *ARPN J. Eng. Appl. Sci.* **2015**, *10*, 5947–5953.
50. Carrington, A.M.; Manuel, D.G.; Fieguth, P.W.; Ramsay, T.; Osmani, V.; Wernly, B.; Bennett, C.; Hawken, S.; Magwood, O.; Sheikh, Y.; et al. Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *45*, 329–341. [CrossRef]
51. Akula, R. Fraud identification of credit card using ML techniques. *Int. J. Comput. Artif. Intell.* **2020**, *1*, 31–33. [CrossRef]