DOI https://doi.org/10.30525/978-9934-26-142-8-2

SYSTEM FOR CLUSTERING AND PREDICTING CUSTOMER BEHAVIOR IN A GIVEN DATA SET

Boyko N. I.

INTRODUCTION

Companies usually have a lot of information about their customers: location, e-mail, contact information. If the user fills out forms - the quality and quantity of this data increases. This can be used to predict the possibility of purchase using the methods of predictive analytics. These predictions can then help the company establish business processes and solve sales department optimization problems. The purpose of this work is to study how machine learning can be used to predict leadscoring[1].

Data from previous periods were used as training data for classification algorithms, and data on purchasing moments were used as limits for contacts who made a purchase in the past. Different ways of aggregating data over time have been considered to ensure that customer constraints do not affect the final model. The performance of the model was evaluated using cross-validation. The results confirm that to assess the possibility of making a purchase by the customer actually using such algorithms of learning with the teacher as randomforest and highlight some important information for the business using visual analytics[2].

1. The problem's prerequisites emergence and the problem's formulation

From the very beginning of its existence, companies from all spheres use data, and always use them for such purposes – to solve existing problems, make decisions that will improve business processes and analyze history[3]. Most of the most important decisions are about attracting customers. For example, which customers should the sales department focus on. During the customer engagement phase, companies try to convert

[1] Benhaddou Y., Leray Ph. Customer relationship management and small data - application of bayesian network elicitation techniques for building a lead scoring model. URL: https://hal.archives-ouvertes.fr/hal-01619307/document.

[2] Dutta G. Lead Scoring with Random Forest. URL: https://www.kaggle.com/gauravduttakiit/lead-scoring-with-random-forest.

[3] Michiels I. Lead Lifecycle Management. URL: https://www.ontargetpartners.com/wp-content/uploads/2010/02/Building_A _Pipeline_That_Never_Leaks_by_AberdeenGroup.pdf.

leads from customers using various methods, such as sending them emails and making phone calls. However, all leads are different, some of them have a much higher probability of becoming a new customer than others, or vice versa. Companies usually do not want to spend their time on unpromising leads, because every extra hour of work of sellers has a certain price. This brings us to the question: how can a company distinguish between “Good” and “Bad” leads?

Moreover, companies want to know the reason why customers have chosen them to improve their marketing decisions in the future. Significant sums are spent each year on advertising and promoting their products or services, but the path of customers still seems unknown to most. Knowing the customer and having an assessment of the possibility of generating these customers from leads, the sales department can greatly improve its work.

The main purpose of this work is to build a system that will solve the problem of converting potential customers into those who use the services of the user company. The task of the final product is to simplify the work of sales departments, improve marketing campaigns, scale existing systems.

2. The analysis of existing methods for solving the problem and formulating a task for the optimal technique development

2.1 Manual evaluation of potential customers

The assessment of potential customers is used by the company's management in determining the priorities that lead to the target. To begin with, companies can count potential customers according to the data they have about them. For example, if a contact visits a website, he is awarded 5 points, but if a contact sends an email, he receives 25 points. The idea here is that sellers should only spend their time on contacts who have a high rating, which, assuming a reliable scoring procedure, means that they will also have a high probability of converting sales. However, there is an obvious problem in assessing people’s behavior, which is based only on the feelings and knowledge of the acquisition process. First, manual customer counting is very imperfect in terms of statistical analysis. Second, assigning the correct value to the activity depends on the availability of data from the company about previous customers for a fairly long period of time. When manually counting leads, points can be assigned based on a set of fixed rules. The table in Figure 1 is formed as follows: the first column Activity contains any user action. In the case of manual assessment of leads, the list of these actions is determined by the sales team. Having the first column -

the second is created, where there is an assignment of points to each action. In this case, these points can be awarded by a team of analysts or based on previous marketing campaigns. As you can see in the picture, the user who is responsible for clicking and filling in the feedback form received the most points. Giving it the highest number of points means that this action is very important and shows a person's interest in the product.

Activity	Points
Form/Landing Page Submission	+ 5
Submitted "Contact Me" Form	+25
Received an Email	0
Email Open	+1
Email Clickthrough	+3
Registered for Webinar	+3
Attended Webinar	+10
Downloaded a Document	+5
Visited a Landing Page	+2
Unsubscribed from Newsletter	-2
Watched a Demo	+8
Contact is a CXO	+5
Visited Trade Show Booth	+3
Visited Pricing Page	+10

Fig. 1. Example of a matrix for manual calculation in lidogeneration.

The result of such calculations is a matrix of points (Fig. 1). This matrix was taken from the article "Marion, G. 2016. LeadScoringis-Broken"[4]. In it, the author says that after the analysis, he did not find a statistical difference between the conversion of leads, which were marked as those most likely to be converted, and the spontaneous selection of any ice and its conversion. His main idea is that it is impossible for a person without experience in statistics and available data on previous sales, to correctly place the scores in this matrix. He argues that the process of constantly adjusting scores is a very time consuming process, and that the time used can be spent more efficiently elsewhere.

The amount of data needed to create an accurate model is huge. For example, if company inputs that include pre-sales statistics and user information are included in the model, most of the information becomes either unavailable or unnecessary when manually evaluating leads. During the manual evaluation of potential customers, we only have access to the

[4] Marion G. (2016). Lead Scoring is Broken. Here's What to Do Instead. URL: https://medium.com/marketing-on-autopilot/lead-scoring-is-broken-here-s-what-to-do-instead-194a0696b8a3.

information provided by the user here This means that the history of customer behavior and specifics can be completely lost during the generation process. For example, in B2B sales, decisions are usually made by a group of people, and the method of manual evaluation of potential customers involves only one actor.

Companies generate a lot of information on a daily basis, which means they can't rely on feeling or intuition when implementing a solution that generates leads. On the contrary, they should aim for a decision that has been made on the basis of available information, although in reality it is not so easy to do so. The set of information is usually large and difficult for the human mind. However, the computer can cope with this task. This confirms the need to use technology to generate and predict the path of the client, namely – the predicted analytics.

2.2 Predicted analytics

In the case of lead generation, predictive analytics is described as a diverse set of mathematical and statistical techniques for recognizing patterns and predicting them in the future in a data set. When predictive analytics is used to count potential customers, it is part of predictive marketing – a consumer-oriented marketing approach that aims to improve impressions throughout the customer's life cycle. The approach was developed due to the assumption that nowadays customers expect an individual attitude when interacting with the company. This is made possible by new technologies that were previously unavailable to marketers. Another factor that contributes to the success of forecasting marketing is a sharp reduction in the cost of calculations. This is a critical aspect, because the technologies used in forecasting marketing can be quite expensive.

As mentioned earlier, predictive analytics is best described as a set of methods used to obtain statistics. These methods are often presented as mathematical / statistical algorithms, or machine learning algorithms. These algorithms are sorted into three categories: with teacher, without teacher and with reinforcement. Algorithms for learning with teacher evaluate the output of the input data, for example, by estimating the likelihood of customer interaction with the company. Non-teacher learning algorithms try to find certain patterns in data without output information, such as looking at a customer base and trying to find customer groups that are different from each other.

Reinforced algorithms look for hidden patterns in the data to recommend the next best action. It can also be used to recommend products

to customers based on the entire shopping history of the customer base or other preferences.

The purpose of estimating leads is to obtain a value that will describe the probability of conversion of this ice to the customer. In this process, the input data are data from the company, and the output is a value that represents the probability of conversion of ice to the client. Thus, controlled learning algorithms are suitable for this task. Types of lead counting algorithms can also help in more complex lead assessment cases. For example, if someone intends to evaluate lead clients in a client segment, you can first use a non-teacher learning algorithm to create segments.

2.3 Analysis of existing systems

2.3.1 HubSpot

One of the best features of the HubSpot solution for identifying potential customers is the fact that it is already included in one of the most popular marketing automation platforms available on the market today. This solution is available to all enterprise customers, which is great for those who want to have a pleasant experience in a single window. The solution is provided with a default model based on templates used by successful customers, but for those who need it This solution is ideal for those who have already kept involved and unused contacts in HubSpot. The software provided in the application will determine which customers fall into the low, medium or high rating categories. The software even provides a pie chart based on several analytical criteria. (Table 1)

Table 1

Advantages and disadvantages of the HubSpot system

Advantages	Disadvantages
Part of the HubSpot ecosystem	Additional options and functionality can be difficult for ordinary users to learn.
Comes with a pre-designed and customized model, which studied on the data of successful clients..	Smaller companies with fewer leads may not need such a complex solution.
Possibility to set up an automatic e-mail for the sales team when a new potential customer appears.	

2.3.2 Infer

Unlike HubSpot, Infer is a dedicated lead counting platform designed to connect to a CRM or marketing automation solution. The software uses an

active API connection, which allows it to seamlessly connect to virtually any CRM solution that is currently available or will be available.

The software also allows managers to seamlessly use thousands of data points based on firm, technology, or demographic information. The software even has built-in information about 19 million companies and 42 million potential customers. Like the best intelligent software, it will use machine learning to determine patterns in both B2B and customers using data obtained from CRM. (Table 2).

Table 2

Advantages and disadvantages of the Infer system

Advantages	Disadvatages
Automatically throws ratings into CRM or any other system.	Very high price.
Use their modified version of logistic regression, which speeds up the work.	Using the method of logistic regression reduces the accuracy of predictions.
Has the ability to predict the conversion of leads for a given period of time.	

2.3.3 PipeCandy

While solutions like Infer are great for traditional B2B because they use like-minded communities, solutions like PipeCandy work just as well in B2C and e-commerce. As a result, PipeCandy is a great tool for organizations looking to collaborate or sell to other companies in this particular space.

PipeCandy easily integrates with CRM to determine winnings and losses to create new evaluation results for potential customers. Analytical indicators are very clear and contain concisely organized visual material that can be used to adjust your strategy.

PipeCandy works well for companies with smaller datasets, using the "Attribute Importance". This feature allows managers to decide which factors are most valuable when counting leads. For example, if you want to add more value to those potential customers who have higher income, the software allows you to easily change its methodology (Table 3).

Table 3

Advantages and disadvantages of the PipeCandy system

Advantages	Disadvantages

The Attribute Importance feature gives managers a lot of opportunities.	The developed model contains a lot of shortcomings, for example, one of the most famous is the error when the model classifies the company "Apple" as a food supplier.
The company provides several plans with different prices.	The solution comes in a comprehensive package that increases the price and can provide unnecessary and redundant functionality
The solution provides additional options for e-commerce customers and these solutions are quite accurate.	

2.3.4 Maroon.ai

Maroon.ai is intelligent software that not only evaluates potential customers, but also helps to generate new potential customers. The software is designed for what the company calls “deep context discovery”, which is designed to help organizations identify their target customers. This makes the solution suitable for beginners as it automates some key processes.

The software is also great for integrating into existing CRM solutions such as Salesforce and Informatica, and the API is customizable for those who want to integrate an AI-driven system into other products. Maroon has a variable pricing structure that offers a significant number of options – for smaller organizations there is even a free version of Maroon.ai. (Table 4)

Table 4

Advantages and disadvantages of the Maroon.ai system

Advantages	Disadvantages
This is a very accurate decision.	H Not customer-oriented design.
Provides additional opportunities for prioritizing between leads and customers	A complete solution that contains many unnecessary features.
Contains ready integrations with popular CRM platforms.	

Table 5 compares existing solutions and the solution being developed.

Table 5

Compares existing solutions and the solution being developed

Solutions	Price (1-10)	Integration complexity	Additional features (1-10)	Accuracy (1-10)
HubSpot	10	3	10	9
Infer	10	5	9	6
PipeCandy	9	4	7	8
Maroon.ai	7	1	7	7
Solution to be developed	3	4	5	9

It can be concluded that there are currently products on the market that provide accurate and effective predictions with many additional features, but their price is very high. The product being developed, on the other hand, provides a smaller package of features, should not be inferior in accuracy, but comes at a lower price. It is more difficult to integrate, but the time spent initially compensates for the financial costs.

2.4 Description of algorithms to be used

As mentioned in the introductory section, teacher training methods are a category of machine learning methods that can be used to evaluate an established outcome variable. In the case of counting leads, the result to be evaluated is the probability that the contact will acquire the product[5].

Teaching methods with a teacher can be divided into two different categories: classification and regression methods. The difference between the two types is that regression is used to estimate continuous values and classification is used to estimate categorical results. In addition, classification models have two results: one is a value from 0 to 1, which represents the probability that the example belongs to a certain category, and the other is a discrete category.

The evaluation of leads can be considered as a problem of classification.

[5] Dutta G. Scoring with Decision Tree. URL: https://www.kaggle.com/gauravduttakiit/lead-scoring-with-decision-tree/notebook.

2.4.1 Models of classification trees

Classification trees are part of a family of tree-based models. They consist of nested if-then statements derived from variables found in a dataset. One example of a simple classification tree can be seen in Figure 2.

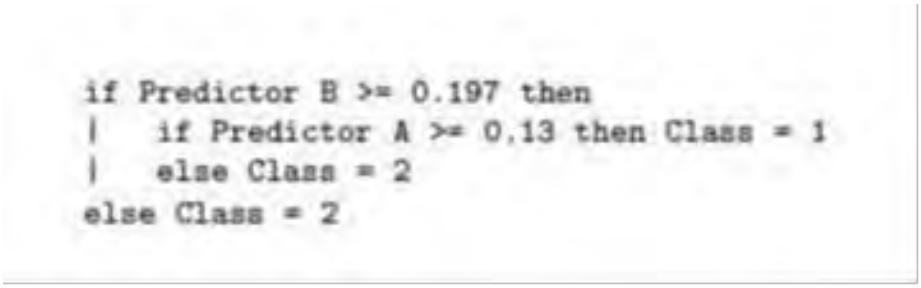

Fig. 2. Example of a simple classification tree

In this figure, you can identify the three end nodes and two branches. End nodes, or leaves, are nodes that lay a tree and contain the result. Possible results in Figure 2 include “class 1” and “class 2”.

The branches shown in Figure 4 are the points where the decision tree makes decisions. The first distribution checks whether predicate B is greater than or equal to 0.197, if so, the decision tree checks whether predicate A is equal to or greater than 0.13. If predicate A is greater than 0.13, then the decision tree will reach the terminal node, which will conclude that the result will be “Class 1”.

2.4.2 Random forest

Algorithm random forest attempts to alleviate the problems of variance or instability of the classification tree algorithm using a modified version of bugging aggregation – a method used to reduce the variance of predicted prediction functions. In essence, a random forest creates several decorelated models of the decision tree and calculates their average value as a basis for predicting the result.

The general algorithm for random forests was created by Breiman in 2001 after evaluating different approaches to introducing randomness into the tree construction process. The reason why the element of randomness was introduced was to increase the efficiency of the decision tree model through decorrelation.

2.4.3 Logistic regression

Logistic regression models assume the probability that the input data belong to a certain class in the binary classification problem (ie, two possible output classes). This means that the result must remain between 0

and 1 for both classes and that their sum must be equal to 1. The result can be interpreted as a probability that estimates the probability of the result[6].

Logistic regression is a simple but popular model that belongs to the family of generalized linear models. It can also be used to asses predicates used in the model, for example, to assess whether a predicate has a statistically significant dependence on the probability of an outcome.

The model is often used in biostatistical programs where there are two possible outcomes or classes. Examples of these types of classification problems include determining whether or not a patient has the disease.

2.5 Building a model

To choose the model that is the most suitable in our case – we use the algorithm of k-fold cross-approval for 10 subsamples. Next, we consider different options for data aggregation to see which method of aggregation gives the lowest error and the highest performance.

Aggregation 1

These results were obtained by setting the last date in the input data for people who will not become buyers, which would fall at the end of the selected period of time. For buyers, the expiration date was set to the value of the time of purchase. As a result, people who do not become buyers will have very different aggregate values depending on when they were active. However, buyers will still have quite similar aggregate values because the date of their last action corresponds to the time of purchase. Of course, some non-buyers have similar aggregate values to buyers. The model is aware of this difference and becomes effective in distinguishing non-buyers from buyers based on variables related to the expiration date. A matrix of discrepancies is used to describe the results of aggregation. This is a table that allows you to assess the accuracy of the algorithm used in teaching with the teacher. In this case, each row represents a sample of the predicted class, and the column represents an existing one.

	true false	true true	class precision
pred. false	8346	74	99.12%
pred. true	1134	676	37.35%
class recall	88.04%	90.13%	

Fig. 3. Random forest

[6] Dutta G. Lead Scoring with XGB Classifier. URL: https://www.kaggle.com/gauravduttakiit/lead-scoring-with-xgbclassifier.

Figure 3 shows the matrix of discrepancies for a random forest. A sample of 10230 values belonging to the classes is given: 1 – lead that will become a customer and 0 – one that will not. Real (valid) sets are placed in a column, while predicted sets are placed in rows. Having these two sets, a matrix of discrepancies is created, which summarizes these results of the classifier check. It can be concluded that for a random forest, the accuracy of class prediction is 99.12% and 37.35%, respectively.

	true false	true true	class precision
pred. false	8395	64	99.24%
pred. true	1085	686	38.74%
class recall	88.55%	91.47%	

Fig. 4. LightGBM

Figure 4 shows the mismatch matrix for LightGBM. A sample of 10230 values belonging to the classes is given: 1 - lead that will become a customer and 0 - one that will not. Real (valid) sets are placed in a column, while predicted sets are placed in rows. Having these two sets, a matrix of discrepancies is created, which summarizes these results of the classifier check. It can be concluded that for LightGBM the accuracy of class prediction is 99.24% and 38.74%, respectively.

	true false	true true	class precision
pred. false	8275	92	98.90%
pred. true	1205	658	35.32%
class recall	87.29%	87.73%	

Fig. 5. Logistic Regression

Figure 5 shows a mismatch matrix for logistic regression. A sample of 10230 values belonging to the classes is given: 1 – lead that will become a customer and 0 - one that will not. Real (valid) sets are placed in a column, while predicted sets are placed in rows. Having these two sets, a matrix of discrepancies is created, which summarizes these results of the classifier check. It can be concluded that for logistic regression the accuracy of class prediction is 98.90% and 35.32%, respectively.

	true false	true true	class precision
pred. false	9163	272	97.12%
pred. true	317	478	60.13%
class recall	96.66%	63.73%	

Fig. 6. Decision tree

Figure 6 shows the mismatch matrix for the decision tree. A sample of 10230 values belonging to the classes is given: 1 – lead that will become a customer and 0 - one that will not. Real (valid) sets are placed in a column, while predicted sets are placed in rows. Having these two sets, a matrix of discrepancies is created, which summarizes these results of the classifier check. It can be concluded that for the decision tree the accuracy of class prediction is 97.12% and 60.13%, respectively.

Aggregation 2

This aggregation method sets the completion date for buyers as the last activity before purchase. For non-buyers, the last activity sets the date of the last activity. This method recorded a bias from aggregation method 1, but the values of recall and accuracy fell. A matrix of discrepancies is used to describe the results of aggregation. This is a table that allows you to assess the accuracy of the algorithm used in teaching with the teacher. In this case, each row represents a sample of the predicted class, and the column represents an existing one.

	true false	true true	class precision
pred. false	6551	252	96.30%
pred. true	2929	498	14.53%
class recall	69.10%	66.40%	

Fig. 7. Random forest

Figure 7 shows the matrix of discrepancies for a random forest. A sample of 10230 values belonging to the classes is given: 1 – lead that will become a customer and 0 – one that will not. Real (valid) sets are placed in a column, while predicted sets are placed in rows. Having these two sets, a matrix of discrepancies is created, which summarizes these results of the classifier check. It can be concluded that for a random forest, the accuracy of class prediction is 96.30% and 14.53%, respectively.

	true false	true true	class precision
pred. false	6544	230	96.60%
pred. true	2936	520	15.05%
class recall	69.03%	69.33%	

Fig. 8. LightGBM

Figure 8 shows the mismatch matrix for LightGBM. A sample of 10230 values belonging to the classes is given: 1 – lead that will become a customer and 0 - one that will not. Real (valid) sets are placed in a column, while predicted sets are placed in rows. Having these two sets, a matrix of discrepancies is created, which summarizes these results of the classifier check. We can conclude that for LightGBM the accuracy of class prediction is 96.60% and 15.05%, respectively.

	true false	true true	class precision
pred. false	5492	174	96.93%
pred. true	3988	576	12.62%
class recall	57.93%	76.80%	

Fig. 9. Logistic Regression

Figure 9 shows the mismatch matrix for logistic regression. A sample of 10230 values belonging to the classes is given: 1 – lead that will become a customer and 0 – one that will not. Real (valid) sets are placed in a column, while predicted sets are placed in rows. Having these two sets, a matrix of discrepancies is created, which summarizes these results of the classifier check. It can be concluded that for logistic regression the accuracy of class prediction is 96.93% and 12.62%, respectively.

	true false	true true	class precision
pred. false	8553	478	94.71%
pred. true	927	272	22.69%
class recall	90.22%	36.27%	

Fig. 10. Decision tree

Figure 10 shows the matrix of discrepancies for the decision tree. A sample of 10230 values belonging to the classes is given: 1 – lead that will become a customer and 0 – one that will not. Real (valid) sets are placed in a column, while predicted sets are placed in rows. Having these two sets, a matrix of discrepancies is created, which summarizes these

results of the classifier check. It can be concluded that for the decision tree the accuracy of class prediction is 94.71% and 22.69%, respectively.

Aggregation 3

These results were obtained by determining a random date between the first and last activity for those who will definitely not become buyers. For buyers, the completion date is set as the last action before the decision to purchase the product.

Setting a date from the first to the last activity (not including) will always exclude the last action of users – this can increase the accuracy to identify non-buyers, as they will always have less activity in this environment. However, by choosing a random date between the first and last activity for people who will not become customers, we simulate their activities at different stages of the client's life, which is one of the best ways to teach the model. A matrix of discrepancies is used to describe the results of aggregation. This is a table that allows you to assess the accuracy of the algorithm used in teaching with the teacher. In this case, each row represents a sample of the predicted class, and the column represents an existing one.

	true false	true true	class precision
pred. false	7496	232	97.00%
pred. true	1984	518	20.70%
class recall	79.07%	69.07%	

Fig. 11. Random forest

Figure 11 shows a matrix of inconsistencies for a random forest. A sample of 10230 values belonging to the classes is given: 1 – lead that will become a customer and 0 – one that will not. Real (valid) sets are placed in a column, while predicted sets are placed in rows. Having these two sets, a matrix of discrepancies is created, which summarizes these results of the classifier check. It can be concluded that for a random forest the accuracy of class prediction is 97.00% and 20.70%, respectively.

	true false	true true	class precision
pred. false	7294	163	97.81%
pred. true	2186	587	21.17%
class recall	76.94%	78.27%	

Fig. 12. LightGBM

Figure 12 shows the mismatch matrix for LightGBM. A sample of 10230 values belonging to the classes is given: 1 – lead that will become a

customer and 0 – one that will not. Real (valid) sets are placed in a column, while predicted sets are placed in rows. Having these two sets, a matrix of discrepancies is created, which summarizes these results of the classifier check. We can conclude that for LightGBM the accuracy of class prediction is 97.81% and 21.17%, respectively.

	true false	true true	class precision
pred. false	6066	127	97.95%
pred. true	3414	623	15.43%
class recall	63.99%	83.07%	

Fig. 13. Logistic Regression

Figure 13 shows a mismatch matrix for logistic regression. A sample of 10230 values belonging to the classes is given: 1 – lead that will become a customer and 0 – one that will not. Real (valid) sets are placed in a column, while predicted sets are placed in rows. Having these two sets, a matrix of discrepancies is created, which summarizes these results of the classifier check. It can be concluded that for logistic regression the accuracy of class prediction is 97.95% and 15.43%, respectively.

	true false	true true	class precision
pred. false	8936	477	94.93%
pred. true	544	273	33.41%
class recall	94.26%	36.40%	

Fig. 14. Decision tree

Figure 14 shows a mismatch matrix for the decision tree. A sample of 10230 values belonging to the classes is given: 1 – lead that will become a customer and 0 – one that will not. Real (valid) sets are placed in a column, while predicted sets are placed in rows. Having these two sets, a matrix of discrepancies is created, which summarizes these results of the classifier check. It can be concluded that for the decision tree the accuracy of class prediction is 94.93% and 33.41%, respectively.

Aggregation 4

In this version, for users who do not become customers, a completely random value was set between the start and end (including boundaries) of the activity. For those who become a customer, the date of their purchase was chosen. The results seem biased because buyers use a predetermined expiration date that corresponds to the time of purchase, and non-buyers use a randomly generated date. This means that the last of the non-buyers

will always be left out. In any case, the model notices these subtle differences in the way end dates are affixed. A matrix of discrepancies is used to describe the results of aggregation. This is a table that allows you to assess the accuracy of the algorithm used in teaching with the teacher. In this case, each row represents a sample of the predicted class, and the column represents an existing one.

	true false	true true	class precision
pred. false	8144	132	98.41%
pred. true	1336	618	31.63%
class recall	85.91%	82.40%	

Fig. 15. Random forest

Figure 15 shows the matrix of discrepancies for a random forest. A sample of 10230 values belonging to the classes is given: 1 – lead that will become a customer and 0 – one that will not. Real (valid) sets are placed in a column, while predicted sets are placed in rows. Having these two sets, a matrix of discrepancies is created, which summarizes these results of the classifier check. It can be concluded that for a random forest the accuracy of class prediction is 98.41% and 31.63%, respectively.

	true false	true true	class precision
pred. false	8017	76	99.06%
pred. true	1463	674	31.54%
class recall	84.57%	89.87%	

Fig. 16. LightGBM

Figure 16 shows the matrix of discrepancies for LightGBM. A sample of 10230 values belonging to the classes is given: 1 – lead that will become a customer and 0 – one that will not. Real (valid) sets are placed in a column, while predicted sets are placed in rows. Having these two sets, a matrix of discrepancies is created, which summarizes these results of the classifier check. It can be concluded that for LightGBM the accuracy of class prediction is 99.06% and 31.54%, respectively.

	true false	true true	class precision
pred. false	7814	97	98.77%
pred. true	1666	653	28.16%
class recall	82.43%	87.07%	

Fig. 17. Logistic Regression

Figure 17 shows a mismatch matrix for logistic regression. A sample of 10230 values belonging to the classes is given: 1 – lead that will become a customer and 0 – one that will not. Real (valid) sets are placed in a column, while predicted sets are placed in rows. Having these two sets, a matrix of discrepancies is created, which summarizes these results of the classifier check. It can be concluded that for logistic regression the accuracy of class prediction is 98.77% and 28.16%, respectively.

	true false	true true	class precision
pred. false	9012	324	96.53%
pred. true	468	426	47.65%
class recall	95.06%	56.80%	

Fig. 18. Decision tree

Figure 18 shows the mismatch matrix for the decision tree. A sample of 10230 values belonging to the classes is given: 1 – lead that will become a customer and 0 – one that will not. Real (valid) sets are placed in a column, while predicted sets are placed in rows. Having these two sets, a matrix of discrepancies is created, which summarizes these results of the classifier check. It can be concluded that for the decision tree the accuracy of class prediction is 96.53% and 47.65%, respectively.

2.6 Choice of aggregation model

Table 6

The difference between aggregation models

Aggregation type	End date of customer	End date of non-customer	Deviation	The best model
1	Date of purchase	Last date in data	High	LightGBM (0.955)
2	Last activity before purchase	Last activity	None	LightGBM (0.761)
3	Last activity before purchase	Last activity before the random date	Low	LightGBM (0.843)
4	Date of purchase	Randomly selected date	Medium	LightGBM(0.935)

If we look at the different types of aggregation methods that were discussed in the previous section, we can highlight some due to the magnitude of the error, or data loss due to the fact that classes are interpreted differently. Data loss occurs when there are attributes that

indicate the model to which class this example belongs. This prevents the model from correctly predicting the behavior of potential customers. Table. 6 describes how the aggregation methods differ from each other, shows the values of the error value and the best AUC model (area under the ROC curve). The column with the values of the error value is based on the descriptions of different types of aggregations from the previous paragraph and represents a subjective opinion about these models.

Aggregation method 1 showed a high error rate, method 4 showed a larger error than 2 and 3, while 2 showed no error at all. We choose the method of aggregation 2 as the most effective – no error and high AUC compared to other methods.

2.7 Comparison of models

To compare the models, we use the results from the previous paragraph, where the method of aggregation was chosen. Table 7 shows the performance of each model. The accuracy indicator should not be relied on, as in this case there is a large imbalance in the data. Instead, AUC and Youden are ideal for overall model evaluation. Sensitivity and specificity must also be taken into account to understand whether the model has correctly divided the two classes into categories.

Table 7

Performance of each model

Model	Accuracy	AUC	Sensitivity	Specificity	Youden
Decision tree	86.27%	0.749	36.27%	90.22%	0.265
Random forest	68.91%	0.723	66.40%	69.10%	0.355
Logistic regression	59.32%	0.698	76.80%	57.93%	0.347
LightGBM	69.05%	0.761	69.33%	69.03%	0.384

As expected, decision tree models and random forest are not as efficient as LightGBM. They received less in all categories except specificity. The decision tree that was created has a maximum depth of 10, which means that the maximum number of decision points that a branch can have is 10. Despite the fact that decision trees are known for their high clarity, depth 10 means that there may be 2 ^ 10 possible solutions in the end, it is a large amount of information to process. Even the reduction of the resulting decision tree proved that reducing and reducing the depth of the tree in the future can adversely affect various performance metrics. The random forest model was created using 100 decision trees and has a pretty good score, but I chose the LightGBM algorithm.

Logistic regression has generally been included in this comparison to show what can be achieved using the linear classification method compared to more complex, non-linear machine learning algorithms. However, even with small scores in general, the linear regression algorithm still defeats decision trees in Yougen estimation. The model also has the highest sensitivity and lowest specificity, which means that it is better at defining more important classes than other models, but worse at defining less important classes. As a result of these studies, the LightGBM model was chosen as the main one, which will be considered in the next section.

2.8 Dataset analysis

BankMarketingDataSet dataset has been used during this work.

The data contains direct information from a marketing company conducted by the Portuguese Banking Institute. The purpose of the classification in this dataset is to predict whether the client will sign a deposit agreement (variable y). The marketing campaign was based on phone calls and in most cases more than one call was needed to find out the result.

Input variables:

age – age (number)

job: type of person employment (categories: 'admin.','bluecollar','entrepreneur','housemaid','management','retired','self - employed','services','student','technician','unemployed','unknown')

marital: marital status (categories: 'divorced','married','single', 'unknown'; note: 'divorced' means divorced or widowed).

education: education level (categories: 'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course',' university.degree','unknown')

default: does he have a bank loan? (categories: 'no','yes','unknown')

housing: does he have a home loan? (categories: 'no','yes','unknown')

loan: does he have personal loans? (categories: 'no','yes','unknown')

contact: type of possible communication with the client (categories: 'cellular','telephone')

month: last month of contact with the client (categories: 'jan', 'feb', 'mar', …, 'nov', 'dec')

day*of*week: the last day of the week of contact with the client (categories: 'mon','tue','wed','thu','fri')

duration: duration of the last contact with the client in seconds (numeric).

An important note: this attribute strongly affects on the final value (for example, if duration = 0 then y = 'no'). However, the duration is not known before the call. Once the call has taken place y will obviously be known. Therefore, this attribute should only be included for guidance purposes and discarded if the goal is a realistic forecasting model.

campaign: the number of contacts made with this client within a single campaign (numeric, includeslastcontact)

pdays: the number of days that have elapsed since the last contact with the customer after the previous campaign.(numeric; 999 means that the client has not been contacted before)

previous: the number of contacts made with this client before the current campaign.(numeric)

poutcome: the result of a previous marketing campaign (categories: 'failure','nonexistent','success')

Outputvariable (desiredtarget)

y – will the client sign the contract? (binary: 'yes','no')

Let's analyze the selected dataset. The dataset is divided into training and testing: the size of the training data set - (8238, 21) and testing - (32950, 21). There are a total of 10 numeric and 11 categorical attributes. Let's analyze the numerical data from Table 8.

Table 8

Numeric attributes from the data set

	age	cam-paign	pdays	pre-vious	emp. var.rat	Euri-bor3m	nr.emplo-yed
count	3295 0.00	32950.0 0	32950. 0	32950. 0	32950.00	32950.0 0	32950.00
mean	40. 048	2.567	962.4	0.17	0.07	3.61	5166.9
std	10.44	2.77	187.06	0.49	1.57	1.73	72.2
min	17.0	1	0	0	-3.4	0.63	4963.6
max	98.0	56.0	999.0	7.0	1.4	5.04	5228.1

The table shows the values of mean, std, max, min and number for the main digital attributes. If we deduce certain patterns from the text data, we can draw the following conclusions: clients with values of admin, retired, student, and unemployed respond more often. Marital status in general has little influence on predictions. Illiterate customers have a significantly higher level of response, but few. The average value of contact has a significant effect on predictions - an increase of about 4 times for customers

who were contacted by phone. The days of the week do not affect the result, but there is a small trend with more successful calls in the middle of the week.

If a person has been in contact at least once before, he is more likely to convert. Between the ages of 18 and ~ 62, customers who have been contacted more than ten times are more likely to not respond to a campaign. If you contact your customer less than 10 times, will you be more likely to convert?

Customers who have been in contact with less than 10 times have a 4 times higher chance of being converted. The percentage of customers who were converted in the previous campaign was calculated: the result showed that 14% of customers were contacted from the training dataset, 25% of whom were converted. What is the probability that a person will become a customer of two campaigns in a row?

$$P(y=1 \mid poutcome=success) = \frac{P(y=1, poutcome=success)}{P(poutcome=success)}$$

$$P(y=1,\ poutcome=success \mid poutcome \neq nonexistent) = \frac{P(y=1, poutcome=success)}{P(poutcome=nonexistent)}$$

P (y=1,poutcome=success|poutcome≠nonexistent): 0.16

The result is 11.27% higher than the previous conversion estimate, so customers who have already converted in previous campaigns are more likely to convert again.

The pdays attribute is used as an example of data processing (Table 9).

Table 9

The value of the pdays attribute before processing

count	45211.0000
mean	40.197828
std	100.128746
min	-1.000000
25%	-1.000000
50%	-1.000000
75%	-1.000000
max	-1.000000

As you can easily see, many values of the pdays attribute are -1, indicating whether the previous campaign was run on these values. You must replace all -1 values with NaN if these clients will be used in the

current campaign. If you remove or replace all the customers who participated in the previous campaign, thus removing the value of -1, the output will be completely different results (Table 10).

Table 10

The value of the pdays attribute after processing

count	8257.000000
mean	224.577692
std	115.344035
min	1.000000
25%	133.000000
50%	194.000000
75%	327.000000
max	871.000000

The average and median are now completely different, allowing you to use the client in the new campaign.

One of the influential attributes is education. To analyze this attribute, we'll construct a horizontal graph to estimate the median balance for each value of the education attribute. From Figure 19 we can conclude that the client with only school education has the highest median.

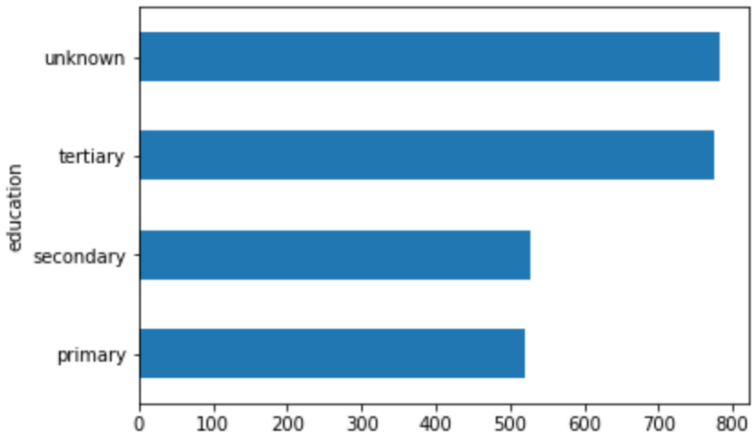

Fig. 19. The median for the education attribute

Let’s consider the general data on the dataset using summary (mydata) str (mydata) (Figure 20).

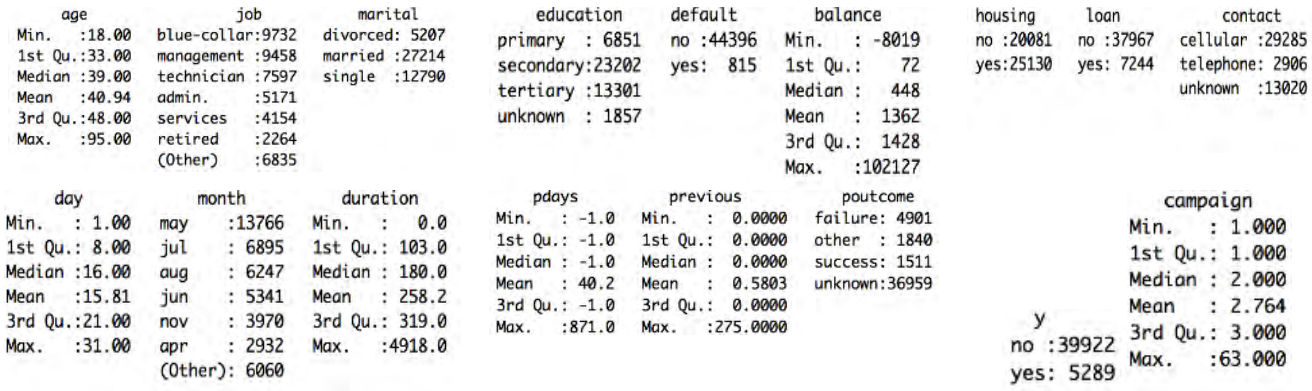

Fig. 20. General information in the dataset

To highlight the impact of attributes, we use permutation-based feature importance. This method means that performance is measured once with and once without resetting trait values. The difference between these two performance values is calculated for each tree and gives the average value, which means the final result. The model can be described by the following formula: PFI_S= E(L($\hat{f}(\widetilde{X_S}, X_C)$,Y))-E(L($\hat{f}$(X), Y)).

2.9 Correction of imbalance

This dataset contains an imbalance of 89:11 (Negative (0): Positive (1)). To correct the imbalance, experiments were performed with the methods RandomUnderSampling, RandomOverSampling and ClassWeightsBalanced.

Class scales directly modify the loss function by giving a larger penalty to classes with larger or smaller weights. The difference in weights will affect the classification of classes during the training phase. The goal is to give penalties to poor classification by putting more weight on the class while reducing its role.

When working with unbalanced data, standard classification indicators do not adequately reflect the effectiveness of the models. For example, let's say a model is created that takes into account the records of the sales department in negotiations with people, and classifies whether they are likely to become a customer.

The most important metrics in such measurements are: Accuracy is defined as the proportion of relevant examples (truepositive) among all examples that are expected to belong to a particular class.

$$\text{Accuracy} = \frac{truepositive}{true\ positive + false\ positive}$$

Precision is defined as the proportion of examples that were assumed to belong to a class, depending on all the examples that actually belong to the class.

$$\text{Precision} = \frac{truepositive}{true\ positive + false\ negative}$$

Next, you can group these two metrics into one and get a variable called f-score and calculated as:

$$F_\beta = (1 + \beta^2)\frac{accuracy * precision}{(\beta^2 * accuracy) + precision}$$

The β parameter allows you to control the difference in importance between accuracy and precision. β <1 focuses more on accuracy, while β> 1 focuses more on accuracy.

Another common tool used to understand model performance is the receiver performance curve (ROC). The ROC curve visualizes the ability of the algorithm to distinguish a positive class from a residual one.

TPR = (true positive)/(true positive+false negative)

FPR = (false positive)/(false positive+true negative)

The RandomUnderSampling and RandomOverSampling methods are methods that are widely used in many models to work with very unbalanced datasets. They consist of removing (under-sampling) members from classes and / or adding them (over-sampling).

Of these three methods, I chose ClassWeightBalanced because it provides the least information loss, has a normal learning time and fits the system solution.

2.10 Results of model training

75.04% of leads received predictions that they will be converted from customers and 29.56% of negative-positive evaluations, 24.96% of negative ones.

After training the model, the following results were obtained:

Quantitative interpretation of ROC (AUC) = 0.8003901184771813

Precision (accuracy / precision) = 0.8003901184771813

Sensitivity (completeness / recall) = 0.6349636070217042

TruePositive Rating: 75.043
FalsePositive rating: 29.566
FalseNegative rating: 24.957.

Based on the ROC-AUC curve and scores, the model assigns the values “high”, “medium”, “low” for each contact in test_set. From the obtained results it is possible to deduce the schedule as in Figure 21. It clearly shows that most of the results were rated medium, which means uncertainty or 50/50. High values are responsible for those records that are most likely to become customers.

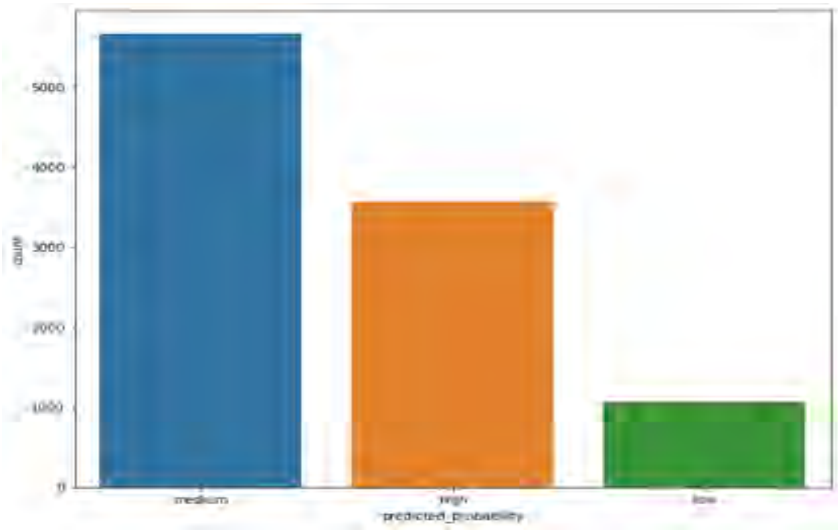

Fig. 21. Developed by DataModel.

2.12 Use of the developed system

The main idea of the developed product is its implementation as a plug-in in popular CRM-systems. In this work, the Salesforce ecosystem was taken as an example. First, a data-model was developed for the sales team.

The Account object represents the company in which all activities are conducted, in this case it is the bank from whose calls the dataset was collected. Contact is an object that contains information about a bank customer. That is, before contacting the client, a contact is created in the database, which contains the necessary information (phone number, address, etc.). The BankClientInfo object has a lookup link to the contact and contains exactly the same information as in the dataset. This object is the most important for the system, because based on its own information, the plugin will determine as much as possible that this person will become a customer. The lead object represents each person who is a participant in

a marketing campaign (Campaign) and only if the developed system gives the result that the person really becomes a customer, the Opportunity object is used, which represents the possibility of selling a product or service (in this case deposit). There are two options for the plugin to work after configuration:

1. Evaluation of individual clients. Basic customer information should be collected before the call. After all the information is present, the sales representative makes a call during which he fills in all the important attributes on the form. Once the call is completed and all the necessary information has been received - press the Calculate button and make the necessary calculations - if the person you contacted is a potential customer - an entry is created for him in the Opportunity object, if not - the status is changed to Failure.

2. If the first option usage provides for action on one record, the second allows such manipulation as MassCRUD. This method is more appropriate to use. His approach is: Let there be a two-week marketing campaign - during this time, the marketing and sales teams communicate with people and collect data, and after receiving all the necessary data at the end of the campaign, a billing process is launched for all customers. An example of the results obtained by this method can be seen in Figure 22 in the poutcome column.

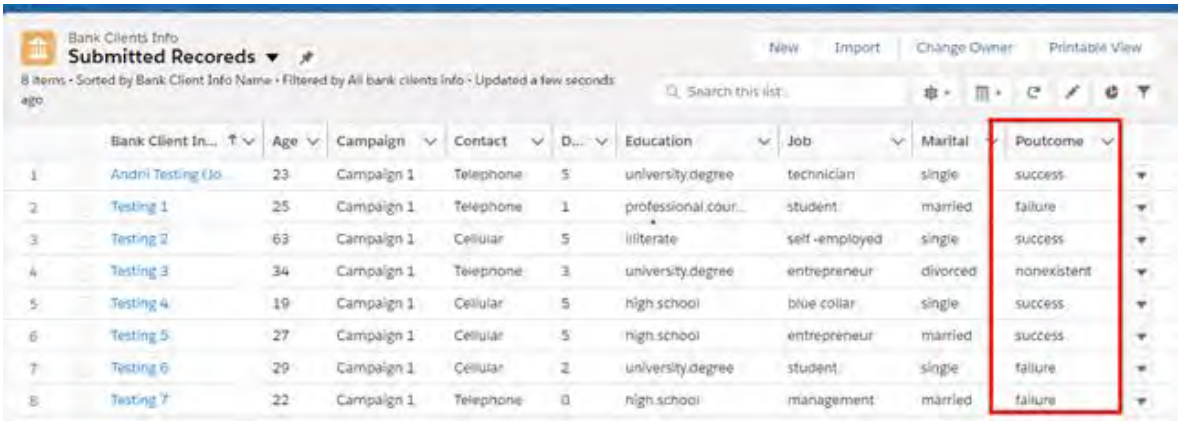

Fig. 22. Lead Client Record Window

The value of Poutcome directly depends on the values returned by the model after training (Fig. 21).

CONCLUSIONS

During this work, the problem of converting leads from customers was considered, the main existing methods were considered, manual and

automatic evaluation of leads was analyzed. In order to develop a model it was necessary to choose a method of machine learning. To make the right decision, we compared methods such as LightGBM, Random Forest, Logistic Regression, Decision Tree, and XGBOOST. After constructing 4 aggregations and subtracting estimates of sensitivity, accuracy, ROC-AUC curve, it was decided to use LightGBM because of better performance, higher learning speed and high efficiency, lower memory usage due to the use of discrete cells and better accuracy than in any other. which amplification algorithm.

A dataset called BankMarketingDataSet was chosen for the training. The data contains direct information from a marketing company conducted by the Portuguese Banking Institute. The purpose of the classification in this dataset was to predict whether the client will sign a deposit agreement (variable y). Feature importance and permutation importance tests were performed and the most important attributes were identified. To correct the imbalance at 89:11 (Negative (0): Positive (1)), experiments were performed with the methods RandomUnderSampling, RandomOverSampling and ClassWeightsBalanced. Of these three methods, I chose ClassWeightBalanced, as it provides the least information loss, has a normal learning time and fits the system solution.

SUMMARY

The results obtained after the work change the approach to the evaluation of potential customers. As described in the analytical review of existing approaches to solving the problem, there are two types of evaluation of potential customers. The first is outdated and less effective - manual evaluation. The proposed solution belongs to the second category – automated evaluation processes. After analysis, it was concluded that the problem with existing systems is their oversaturation. This means that most companies that offer such solutions offer them in comprehensive packages, which include other products, which significantly increases the cost for the end customer. The novelty of the developed product is that it acts as a plug-in to the existing CRM-system of the client, which helps to easily and quickly integrate it. Using the LightGBM algorithm for the leadscoring problem can also be called a new solution, as most existing systems use the random forest method.

REFERENCES

1. Dutta G. Lead Scoring with Random Forest. URL: https://www.kaggle.com/gauravduttakiit/lead-scoring-with-random-forest.

2. Dutta G. Scoring with Decision Tree. URL: https://www.kaggle.com/gauravduttakiit/lead-scoring-with-decision-tree/notebook.

3. Dutta G. Lead Scoring with XGB Classifier. URL: https://www.kaggle.com/gauravduttakiit/lead-scoring-with-xgbclassifier.

4. Marion, G. (2016). Lead Scoring is Broken. Here's What to Do Instead. URL: https://medium.com/marketing-on-autopilot/lead-scoring-is-broken-here-s-what-to-do-instead -194a0696b8a3.

5. Benhaddou Y., Leray Ph. Customer relationship management and small data - application of bayesian network elicitation techniques for building a lead scoring model. URL: https://hal.archives-ouvertes.fr/hal-01619307/document.

6. Michiels I. Lead Lifecycle Management. URL: https://www.ontargetpartners.com/wp-content/uploads/2010/02/Building_A _Pipeline_That_Never_Leaks_by_AberdeenGroup.pdf.

Information about the author:
Boyko Nataliya Ivanivna,
Candidate of Economical Sciences,
Associate Professor at the Department of Artificial Intelligence
Lviv Polytechnic National University
5, Prince Roman street, Lviv, 79000, Ukraine