Application of Data Analytics Techniques for Predicting Customer Churn

Anna Biceková, Oliver Lohaj, František Babič, Marek Puškáš

Department of Cybernetics and Artificial Intelligence, Faculty of Electrical Engineering and Informatics, Technical University of Košice, Letná 1/9, 040 01, Slovakia, anna.bicekova@tuke.sk, oliver.lohaj@tuke.sk, frantisek.babic@tuke.sk, marek.puskas@student.tuke.sk

Abstract: This paper deals with the classification of bank customers based on their characteristics. The aim of this research is to provide an optimal model capable of identifying customers who plan to leave the bank in the near future and switch to competitors. Identifying the customers is crucial because it allows you to allocate funds to marketing more efficiently and ensure competitiveness. As part of the paper, the data were first adequately pre-processed, which was followed by the compilation of eight classification models. These models were then divided into two groups according to the approach to solving class imbalances in the target attribute.

Keywords: customer churn; classification; class imbalance; competitiveness

1 Introduction

Banking is an industry constantly influenced by various factors, such as technological progress, changes in legislation, economic cycles or changing customer behavior. These and other factors can fundamentally affect the position of banks in the market. A key aspect for banks is to monitor the behavior of their existing and potential customers. To ensure competitiveness, it is essential for banks to understand the needs of their clients and identify customer groups with a high risk of exiting. The issue of customer churn from banks has been a long-discussed topic among experts, especially in the context of a dynamic market and a large number of providers of similar services. Due to the constantly changing conditions, banks are forced to monitor current trends and customer requirements in order to be able to respond flexibly to changes in the behavior of their current and potential clients. With the onset of rapid technological progress, the demands of younger generations for personalized and innovative approaches are increasing, which is also reflected in their behaviour when choosing banking services. Unlike previous generations, younger customers have access to more information, which

influences their decision making. Meeting the needs of these customers is thus a challenge for banking service providers, who have to come up with innovative and attractive offers. It is common knowledge that it is financially more difficult for a bank to acquire a new client than to retain one. Acquiring a new customer requires a significant investment in marketing and promotion. According to Ljubičić and Merćep [1], the cost of acquiring a new client in the banking sector is higher than retaining existing clients, with banks investing about $200 to acquire one new client. For this reason, a system that can effectively predict customer churn in the shortest possible time is critical for banks, which could bring significant savings in funds and at the same time increase their competitive advantage. Early identification of a possible departure of a certain group of customers allows banks to adapt and optimize their services to better reflect the needs and preferences of clients.

According to the publication [2] these include the key factors influencing the choice of bank by clients: account maintenance without fees, the amount of bank fees, the quality of electronic banking, free withdrawals from ATMs of your own bank, the availability of branches, advantageous appreciation of funds, references of the bank and the possibility of obtaining several financial products in one place (e.g. a combination of mortgage, pension savings and household insurance). It follows from the above that customer churn is an integral part of the current competitive banking environment. Banks are facing the challenges of a rapidly changing world and the growing demands of clients. Therefore, it is crucial that they invest in personalized and innovative solutions that will allow them to retain existing customers and acquire new ones. It is equally important to implement systems to predict client churn, which can lead to cost optimization and provide banks with a competitive advantage.

2 Related Work

The authors Rahman and Kumar focused on predicting customer churn from commercial banks using data mining in their study from Central University of Kerala [3]. The research used data from a sample of 10.000 customers available on Kaggle, which included attributes such as age, gender, credit score, account balance and number of banking products. Of the total number of customers, 7963 kept the account, while 2037 canceled it. The authors modified the data using mRMR (Minimum Redundancy Maximum Relevance) and Relief algorithms to select the most relevant attributes. Minimum Redundancy Maximum Relevance (mRMR) is a feature selection method that focuses on selecting a subset of relevant features from the original set while trying to minimize the redundancy between the selected features. mRMR identified product count, activity, age, gender, and account balance as key, while Relief added credit card ownership.

Due to the imbalance of the data, the authors used the oversampling method. From all the machine learning models compared, Random Forest (RF) achieved the best results with an accuracy of 95.74% after oversampling. The mRMR and Relief algorithms did not have a significant impact on the accuracy of Decision Tree, which was around 91%. However, the SVM method showed reduced accuracy after oversampling, decreasing from 79.63% to 70.36%.

According to the authors Bansal, Singh et al. [4], five freely available datasets were analyzed, containing over 50.000 records with attributes such as credit score, geography, salary, age, gender, and education. The aim was to predict the customer's bank account cancellation (1) or not (0). They applied both homogeneous (Bagging, Random Forest, Boosting) and heterogeneous (Voting, Stacking) methods of ensemble learning. They evaluated the results using metrics such as recall, precision, geometric mean, and F1 score. The data was pre-processed before analysis, with the missing values supplemented with aggregated data. The datasets were then split 80:20 for training and testing, and further balanced using undersampling and oversampling. Each method was applied to each of the five sets, resulting in 50 results. The average score showed that Random Forest (RF) and HistGradientBoosting (HGB) were the most effective of the collective learning methods, with results of 85.24% and 85.22% respectively within the geometric mean.

The authors Saxena, Singh et al. [5] discussed using a dataset from Kaggle with 10.000 records, containing attributes such as age, gender, credit score, number of bank products, account duration, and a customer churn indicator. Unnecessary attributes, such as line number and nationality have been removed in the preprocessing. The data was divided into five parts using the GMM (Gaussian Mixture Model), with each part being analyzed using three classical machine learning methods (CNN, DT-MR, ANN) and the GMM-ASVM (Gaussian Mixture Model Clustering-based Adaptive Support Vector Machine) technique proposed by the authors. GMM-ASVM achieved the best results, with the success rate of the prediction growing from 89.92% in the first sample to 98% in the fifth. Accuracy ranged from 90.35% to 97.22%, and return from 88.25% to 94.24%. Of the classical methods, ANN had the best results, but the difference from GMM-ASVM was from 4% to 8%.

Tran, Le, and Nguyen focused on two main objectives [6] examining the impact of customer segmentation on the accuracy of customer leakage predictions in the banking sector and evaluating the effectiveness of different machine learning approaches for this prediction. The study used a dataset from Kaggle with 10.127 records and 23 attributes, with two attributes removed for segmentation. Attributes such as age, gender, account status, income category and length of existence of the account remained. When preprocessing the data, the authors applied standardization and normalization, as well as binarization for attributes with two values. Using the K-means method and the Elbow technique, the authors determined the optimal number of clusters is 6. Dataset was split 70:30 for

training and testing set. Machine learning methods, like kNN (k-Nearest Neighbor), LR (Logistic Rregression), DT (Decision Tree) and RF (Random Forest) were used for the analysis, with the results evaluated with metrics such as accuracy, precision, recall and F1 score. The results showed that the RF model achieved the highest success rate (97% cluster average, 97.4% original set). The LR model had the worst results in terms of success rate (87.57%), while KNN achieved the best values in return (98.81%) and RF had the highest F1-score (97%).

According to the authors [7], the study was based on the same Kaggle dataset as several other mentioned works (10,000 records of bank customers). However, it differs in several key aspects of methodology and interpretation of results. Unlike studies that applied advanced approaches such as GMM-ASVM or segmentation using K-means, this work focuses on comparing traditional machine learning algorithms. Interestingly, the authors did not apply complex feature selection methods (such as mRMR or Relief), but instead concentrated on thorough data preprocessing, removal of irrelevant attributes, and class balancing using oversampling. This simplicity in approach enabled a direct comparison of algorithms like Logistic Regression, Naive Bayes, Decision Tree, Random Forest, Support Vector Machine, and k-Nearest Neighbors without modifying the dataset structure. In terms of results, the study confirmed the dominance of the Random Forest model (86% accuracy), while highlighting the relatively low performance of the Naive Bayes model (70%), a more noticeable gap than some other studies. The study thus emphasizes that even with the same dataset, simple approaches and basic models can deliver solid results – especially with proper data preprocessing.

The authors [8] focused their study on predicting customer churn in banks using deep learning techniques, which set them apart from others who primarily relied on traditional machine learning algorithms. Although they used the same Kaggle dataset with 10,000 records, their approach differed mainly in the choice of models and data processing methods. As part of the preprocessing, they applied Min-Max normalization and encoded categorical variables (such as gender and geographic location). They did not use any techniques to balance the imbalanced dataset, which highlights the ability of their models to perform effectively even with skewed data. The main contribution of the study was the implementation of BiLSTM and the hybrid CNN-BiLSTM models. The CNN-BiLSTM model achieved the best results, with an accuracy of 90.2% and the highest F1 score of 0.88, outperforming traditional models such as SVM, logistic regression, and k-NN. Unlike previous works that focused on ensemble methods or feature selection, the authors demonstrated that deep learning can better capture complex patterns in the data and improve prediction performance without the need for additional adjustments to the dataset structure.

The basic idea behind these studies was to predict whether a customer would leave the bank or stay, while data processing methods included standardization, normalization, and techniques to address data imbalances, such as oversampling

and undersampling. One of the common features of all the studies is the wide use of the RF model, which in all cases proved to be a highly accurate model with a success rate of up to 97%. RF has often been compared to other models such as SVM (Support Vector Machine), LR and k-NN. Of these models, RF achieved the best results, while LR and SVM showed lower accuracy, especially when using oversampling.

The differences between the studies lie in the approaches to data segmentation and classification. For example, one study used the Gaussian Mixture Model (GMM) to divide the data into five parts, resulting in increased accuracy of predictions, while other studies used collective learning methods such as Bagging and Boosting. Another study used the K-means method to segment customers and examined how segmentation affects the accuracy of prediction.

The advantages of these studies are in their ability to significantly increase the accuracy of predictions with advanced data mining methods and model optimization. Random Forest excelled in every study where it was used, combining it with techniques such as oversampling or data clustering to provide the best results. The disadvantages mainly consisted of a decrease in the accuracy of models such as SVM and Logistic Regression, especially after oversampling, suggesting the need for additional techniques to manage data imbalances.

3 Methods

Decision tree [9] is one of the most widely used models in machine learning. The basis is to divide data into smaller subsets based on specific rules derived from dataset values. The model has a tree-like hierarchical structure that includes a root node, branches, inner and leaf nodes. The development process begins with choosing the attribute that best divides the data, often based on information gain, which is an entropy-based criterion. This process is repeated until a sufficiently low entropy or other termination criterion is reached.

Random Forest model [10] belongs to the group of ensemble models, which means that it combines the results of several individual models to improve performance. During training, it creates multiple decision trees, and the class selected by the majority of trees is selected as a result in classification. The number of DTs is determined by the user, with each tree trained on a randomly selected subset of training data. For each distribution, a random subset of attributes is selected, and from these, the one attribute with the lowest entropy is chosen. The final result is based on the voting of all trees, with the class with the highest number of votes being the final output.

Logistic regression is a statistical model [11] used for classification that estimates the probability with which a certain element belongs to a particular class. This

model is a form of regression analysis and focuses mainly on binary classification. Its essence is the modeling of the logarithmic function of odds (logit), which is a linear combination of independent variables. The method uses maximum probability (MLE) to estimate parameters that increase the probability of observed data. In addition, this version of logistic regression uses cross-validation to automatically determine the optimal value of the regularization parameter. This model, together with previously mentioned Random Forest reached the highest scores of accuracy in study [12].

A multilayer perceptron (MLP) [13] is a feedforward neural network model where information travels unidirectionally from the input layer through the hidden layer(s) to the output layer. Each connection between neurons is assigned a weight that adjusts during training. The input layer receives the initial data, while the number of neurons in it depends on the dimensionality of the input. Hidden layers process outputs from previous layers, with the number of hidden layers and neurons being adjustable hyperparameters. Bias neurons help modify the activation function, which introduces nonlinearity and allows the model to capture complex patterns in the data. There are neurons in the output layer that determine the final output of the model, for example, the probability in the case of binary classification.

Easy Ensemble Classifier model [14] belongs to the group of ensemble models and is commonly used for classification in cases where there is a significant disproportion between classes in the target attribute. Ensemble models work on the principle of combining the results of several classifiers, which increases the success of the classification. Easy Ensemble Classifier creates multiple balanced subsets by randomly subsampling the majority class and trains the selected classifier on each subset. The final grade for the record is determined by a majority vote of the classifiers. Important hyperparameters include the selection of the estimator, the number of classifiers (n_estimators), the subsampling strategy (sampling_strategy) and the possibility of subsampling with or without replacement.

Bagging Classifier belongs to ensemble models and improves the accuracy and stability of simpler classifiers by training the selected classifier in parallel on multiple random subsets of data. First, so-called "bootstrap samples" are created from the original dataset by randomly selecting repeating records, which means that one record can appear more than once in the sample. On each bootstrap sample, the chosen classifier (e.g., decision tree) is trained independently. The resulting predictions are combined by voting to determine the final prediction. Important hyperparameters include the chosen estimator, the number of classifiers (n_estimators), whether samples should be selected with repetition (bootstrap), and the number of samples for training (max_samples).

Balanced Bagging Classifier is a special type of Bagging Classifier that focuses on solving data imbalances. First, it balances the training data by subsampling the

majority class and/or oversampling the minority class, ensuring a more even representation of the classes. Bootstrap samples are then created from a balanced dataset, with each sample containing approximately the same number of examples from each class. On these specimens, the selected classifier is independently trained. The results are then combined, and the final class is determined by a majority vote. The parameters of this model are similar to those of the Bagging Classifier.

The Adaptive Boosting Classifier is an ensemble model that creates a strong classifier by combining multiple weaker classifiers. During training, individual classifiers adapt and focus on cases that are difficult to classify correctly. At the beginning, all examples in the training dataset are assigned the same weights. Classifiers are trained sequentially, each of them trying to correct the mistakes of the previous one. After each cycle, case weights are adjusted so that more difficult to classify cases are given higher weight. The resulting model is a combination of all classifiers, each having a scale according to its accuracy. Important hyperparameters are: the chosen classifier (estimator), the maximum number of classifiers (n_estimators), and the weight applied to each classifier during the iterations (learning rate) [15].

ADASYN [16] is a classification algorithm designed to balance the uneven representation of classes, especially when improving the performance of models in minor class classification. It works by first determining the degree of imbalance between classes, that is, the ratio of sizes. It then assigns higher weights to those cases from the minority class that are more difficult to classify correctly. Based on this, it generates new synthetic data that are similar to the original data from the minority class but are not identical. The advantage of ADASYN is its ability to create boundary cases that help to better differentiate classes and increase classification accuracy.

3.1 Business Understanding

The aim of this research is to reduce banks' costs associated with customer churn by investing in retaining those clients who are more likely to churn. As we have mentioned, from a financial point of view, it is more difficult for banks to acquire a new customer than to retain them. From the point of view of data mining, the aim is to identify a model(s) that can effectively predict churn based on customer characteristics, thus enabling the identification of customer groups that the bank can target in a best possible manner. This type of task belongs to the field of binary classification, where customers are divided into two classes.

3.2 Data Understanding and Pre-Processing

The dataset we used contains 12 attributes and 10.000 customer records of banks from Germany, France and Spain. Attributes describe various characteristics of customers in relation to the bank, such as the customer's estimated salary, the customer's account balance, credit score, gender, length of time the bank account has existed, whether the customer is considered active and transacts frequently, or credit card ownership. The last attribute churn classifies customers into two classes, i.e. whether the customer left the bank (1) or did not leave (0). The data was divided into a training and test subset in a 70:30 ratio, with the models trained on the training subset and then evaluated on the test subset. To evaluate the success of customer classification, the following metrics were chosen: ROC curve, AUC (area under the ROC curve) and substitution matrix. In Figure 1 is a correlation matrix that shows the interdependencies of the numerical attributes of a dataset. The correlation of attributes with each other is generally low, with the strongest correlation between the account balance and the number of products attributes. However, this interdependence of the mentioned attributes is low enough to be able to keep these attributes in the dataset and provide them to the models in the training subset, without the risk of significant distortion of the results. The dataset originally contained 12 attributes. The 'customer ID' attribute was removed. The correlation matrix shows correlations only between the 6 numerical attributes remaining after the ID was removed, not between all 11 used attributes (6 numerical + 5 nominal/binary). The claim that attribute selection was not necessary refers to the remaining 11 attributes (after removing the ID), from which no further attributes needed to be excluded based on correlation or missing values.

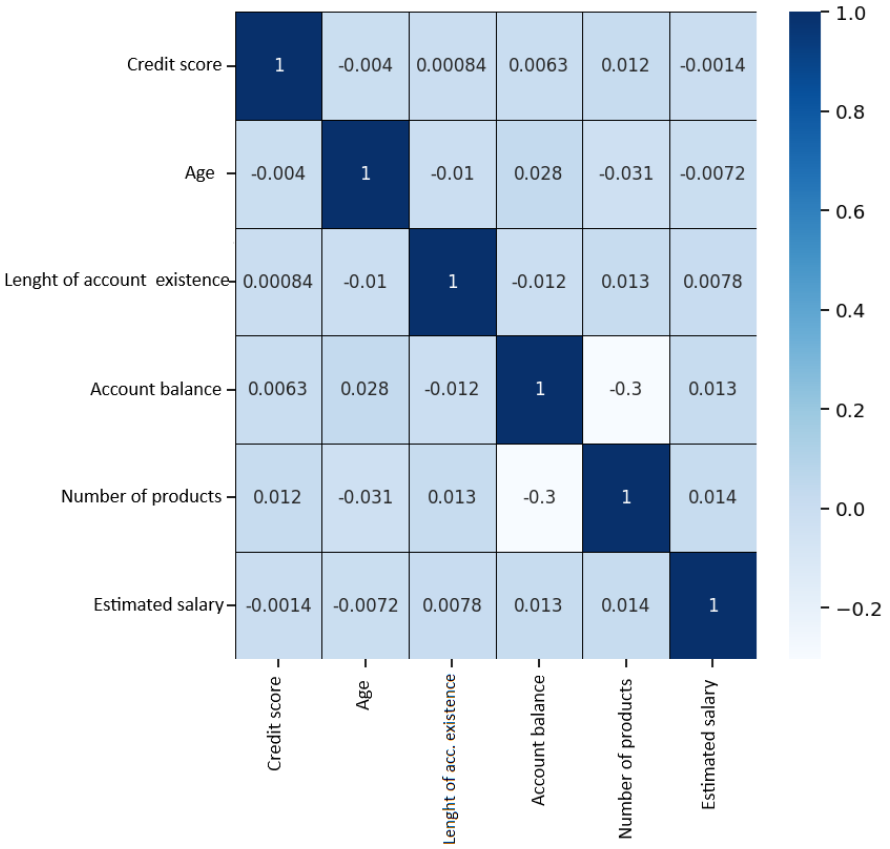

Figure 1
Correlation matrix

As part of the data preparation, the text values in the country and gender attributes were converted to numbers using python and the LabelEncoder function from the preprocessing module, which is part of the scikit-learn library. These changes were made due to further processing and description of the data, as the selected models only support numeric attributes. From the same module, the MinMaxScaler function was also used for the account balance and estimated salary attributes. The values in these attributes were on a wide interval, so they were converted to the range [0,1]. Additionally, the customer ID attribute has been removed because this attribute only corresponds to the ID number that has been assigned to the customer. It has no influence on the customer's decision whether to leave the bank or not. Furthermore, it was not necessary to apply any other methods of data preprocessing, because the dataset selected for analysis did not require them. The dataset contains no missing values and the correlation between attributes is low, so it was not necessary to perform attribute selection.

3.3 Modelling and Evaluation

In this section, we focus on the process of selecting, training and evaluating various classification models aimed at identifying customers with an increased probability of leaving the bank. In this context, eight models were used (see Table 1), which were divided into two categories according to the approach to solving the problem of class imbalance, which is a common phenomenon in predictive tasks, where one class (e.g. outbound customers) is significantly less numerous than the other (remaining customers).

Table 1
Used model

Model Group	
Not adapted to class imbalances	**Adapted to class imbalances**
Random Forest	Easy Ensemble Classifier
Decision tree	Balanced Bagging Classifier
Logistic regression	Bagging Classifier
Multi-layer Perceptron Classifier	Adaptive Boosting Classifier

The first group included modified models, while the problem of class imbalance was solved by using minority class oversampling. For this purpose, the ADASYN (Adaptive Synthetic Sampling) method was used, which generates new synthetic samples for the minority class and thus increases its representation in the data. Class 0 contained 7,963 records and Class 1 contained only 2,037 records. This means the ratio was 20.4% of customers labeled as churned and 79.6% as retained. This approach aims to increase the likelihood of correctly classifying customer churn cases, making models more sensitive to identifying risky customers. The hyperparameter settings were as follows: sampling_strategy='minority', random_state=1, n_neighbors=7. From this we can see that if sampling_strategy='minority' was set, it tells us that synthetic samples will be generated only for the minority class, not for all classes. The random_state=1 parameter sets a random seed, which ensures reproducibility of the results. The last parameter n_neighbors=7 indicates that 7 nearest neighbors were used to generate new records (using k-NN). ADASYN generates synthetic samples near complex decision boundaries, which can introduce noise and reduce class separability. Models like DT, RF, and RL may overfit to these synthetic instances, leading to poorer generalization and lower overall accuracy, even though recall for the minority class improves.

On the other hand, the customized models in the second group worked with unbalanced data without the need for further modification, as they are designed to solve this problem directly. These models included, for example, Balanced Bagging and Adaptive Boosting. This approach avoids oversampling, reducing the risk of over-adapting models to specific cases in training data.

To evaluate the effectiveness of individual models, we used standard performance metrics, which included:

- ROC curve (Receiver Operating Characteristic Curve),
- AUC (Area Under the ROC Curve) to determine the ability of models to correctly distinguish between outbound and remaining customers,
- Confusion Matrix for detailed analysis of prediction errors,
- Accuracy, precision, recall, and F1-score as measures of overall and balanced performance.

The use of these metrics made it possible to evaluate the models' ability to accurately classify customers with a high probability of churn while striking an appropriate balance between sensitivity (recall) and specificity [17].

In the evaluation phase, we first looked at the first subgroup – models with oversampling of the minority class: The models in this group were tested using the ADASYN method to increase the representation of customers who plan to leave the bank (minority class). For some models, this approach was successful, while for others the benefit was not so significant.

1. Random Forest achieved a customer classification accuracy rate of 85.93% before oversampling, but it decreased to 79.43% after using ADASYN. Similarly, the precision of the model has deteriorated from 83.48% to 69.86%.
2. Decision Tree showed an accuracy rate of 85.67% before oversampling, but after ADASYN, this value dropped to 76.33%. The precision of the model has also been reduced from 83.09% to 67.5%.
3. Logistic regression, which achieved an accuracy rate of 80.93% before oversampling, showed a decrease to 72.13% after using ADASYN. However, the recall on the model has improved, from 56.11% to 67.4%.

In the evaluation phase of the second group of models, we worked with models that were not adapted to work with unbalanced data. Models in this group that did not use oversampling showed a better balance between sensitivity and specificity.

1. The Balanced Bagging Classifier (BBC) achieved the highest AUC value of 76.8% and a recall of 76.8%, making it one of the best models in terms of overall performance in this group.
2. The Easy Ensemble Classifier (EEC) was very similar with an AUC of 76.65%, achieving an accuracy rate of 78.43%.
3. The Bagging Classifier (BC) had a high accuracy rate of 83.54% and a precision rate of 85.73%, making it a powerful model, especially when it comes to correctly recognizing customers who stay.

The results showed that non-oversampled models, in particular Balanced Bagging and Easy Ensemble, showed better results in classifying customers who leave the bank, while maintaining an appropriate balance between sensitivity and specificity.

3.4 Deployment

The True Positive (TP) case from the confusion matrix means that the model correctly classified the customer into a class that describes the outbound customers. The True Negative (TN) case means that the model correctly classified the customer into a class that describes customers who did not leave. In the case of false positives (FPs), the costs incurred for customers who were incorrectly identified as outbound were unjustified from the point of view of business objectives, as these customers actually remained in the bank. However, these costs do not represent as high a loss as the customer's departure. A false negative (FN) case describes a situation where the model assessed that the customer did not leave the bank, but in fact left it. That means that in those cases there should have been costs incurred to prevent customers from leaving since the negative consequences of the customer's departure itself are more serious in terms of financial costs than the effort to prevent that situation. As the business goal emphasizes the identification of class 1 customers, the case of FN is worthy of higher attention. Therefore, this case was resulting from the resulting matrices, the assigned weights, which are shown in Table 3.

Table 2
Weight distribution

Case from the resulting matrixes				
	TP	TN	FP	FN
Weight	0	5	-5	-10

When evaluating the classification models, we considered different types of errors and correct predictions so that the model reflects the real financial impact of decisions. A key tool for achieving this goal was the determination of weights for the individual classifications: True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). These weights serve as a form of penalty or reward that influences the optimization of the model according to the set priorities and needs of the organization.

A weight of -€5 was set for false positive cases (FP), i.e. customers who were wrongly marked as leaving. This penalty considers the costs associated with unnecessary marketing activity aimed at a customer without real intention of leaving. For false negative cases (FN), representing the biggest financial loss, the weight was even higher, namely -€10. Indeed, this type of error means that the customers who decided to leave were not correctly identified, and therefore the

bank could not take any steps to retain them, leading to a loss of their future revenues.

Correct classifications were assigned different weights. In the case of True Positives (TP), i.e., customers that the bank already knows have left, the weight was set to €0. This case requires no additional financial investment because these customers are already lost. On the contrary, for True Negatives (TN), i.e., customers who were correctly identified as not leaving, a positive weight of €5 was assigned. This weight considers the savings made by not spending unnecessary marketing costs.

This approach allows the model to focus on minimizing financial losses due to misclassifications (FP and FN) while rewarding correct predictions (TN) that lead to savings.

Table 3
Weight sum results for each model

The first group of models before ADASYN		The first group of models after ADASYN		Second group of models	
Model	Score	Model	Score	Model	Score
RF	7725	RF	5775	EEC	5475
DT	7645	DT	4845	BBC	5865
LR	6225	LR	3585	BC	7665
MLP	6445	MLP	5175	ABC	6535

The final score of each model was calculated using the formula:

$$W_m = \sum_{i=1} p_i \times w_i \tag{1}$$

where W_m is the final score of the model, p_i is the specific case from the pronoun matrix (TP, TN, FP, FN), and w_i is the weight assigned to that case from the pronoun matrix. From the point of view of the business goal and set priorities, the most suitable model is from the first group, namely RF before oversampling, because this model achieved the highest score. Despite the fact that the most effective model in terms of AUC and recall metrics is the BBC model (76.8% for both AUC and recall), from the point of view of the business goal and the fact that we decided to pay the most attention to the case where the customer was incorrectly classified as the one who does not leave (FN), RF turns out to be the most suitable model before using the ADASYN algorithm.

Conclusion

This research aimed to minimize the costs associated with customer churn in banking by identifying high-risk clients who could be targeted more effectively through focused marketing efforts. In today's competitive market, retaining existing customers is more cost-effective and strategic than reacquiring lost ones.

To achieve this objective, a series of classification models were developed and evaluated with respect to the class imbalance of the target variable. The first set of models, including Random Forest and Logistic Regression, was trained using oversampled data, enabling better identification of customers with a high likelihood of churn. The second set of models, specifically designed to handle imbalanced data—such as Balanced Bagging and Easy Ensemble Classifier—was trained without modifying the dataset. The findings revealed that the Balanced Bagging and Easy Ensemble models consistently delivered high performance and robustness without requiring additional data preprocessing, underscoring their suitability for predictive tasks involving unbalanced datasets. This research highlights the true value of data analytics in identifying at-risk customers and offers actionable insights for banks to optimize their marketing resource allocation. The results provide a foundation for further research and practical applications, enabling banks to gain a competitive edge while reducing the financial impact of customer churn.

Discussion

Churn prediction models have important business and ethical implications. On the business side, accurate predictions enable banks to proactively retain customers, potentially reducing financial losses. However, ethical concerns arise regarding fairness, transparency, and data privacy. For instance, targeting only high-risk individuals could unintentionally reinforce biases or lead to unfair treatment. It is essential that such models are used responsibly, with transparent logic and safeguards to ensure compliance with data protection regulations and fairness in decision making. While our models, particularly decision trees and logistic regression, are inherently interpretable and allow for clear decision making based on feature importance, more complex models like ensemble methods (e.g., Balanced Bagging and EasyEnsemble) can act as "black boxes." For real-world applications in banking, where transparency is essential for trust and regulatory purposes, future work could incorporate explainability tools like SHAP or LIME. These methods can help explain complex model decisions, enhance interpretability and support data-driven business decisions.

Acknowledgement

This work was partially supported by Scientific Grant Agency of the Ministry of Education, Research, Development and Youth of the Slovak Republic and the Slovak Academy of Sciences under grants No. 1/0685/21 and 1/0259/24.

References

[1] Ljubičić, K., Merćep, A., Kostanjčar, Z., Churn Prediction Methods Based on Mutual Customer Interdependence, https://doi.org/10.1016/j.jocs.2022.101940

[2] Feng L., Hui L., Meiqian H., Kangle C., Mehdi D., Customer satisfaction with bank services: The role of cloud services, security, e-learning and

service quality, Technology in Society, Volume 64, 2021, 101487, ISSN 0160-791X, https://doi.org/10.1016/j.techsoc.2020.101487

[3] Rahman, M., Kumar, V., Machine Learning Based Customer Churn Prediction in Banking, doi: 10.1109/ICECA49313.2020.9297529

[4] Bansal, A., Singh, S., Jain, Y., Verma, A., Analysis of Ensemble Classifiers for Bank Churn Prediction, doi: 10.1109/ICCCIS56430.2022.10037623

[5] Saxena, A., Singh, A., Govindaraj, M., Analyzing Customer Churn in Banking: A Data Mining Framework, doi: 10.31893/multiscience.2023ss0310

[6] Tran, H., D., Le, N., Nguyen, V., H., Customer Churn Prediction in the Banking Sector Using Machine Learning-Based Classification Models, doi: 10.28945/5086

[7] Ahmed, Md P., et al. "A Comparative Study of Machine Learning Models for Predicting Customer Churn in Retail Banking: Insights From Logistic Regression, Random Forest, GBM, and SVM." Journal of Computer Science and Technology Studies, Vol. 6, No. 4, 8 Oct. 2024, pp. 92-101, doi:10.32996/jbms.2024.6.4.12

[8] A. Manzoor, M. Atif Qureshi, E. Kidney and L. Longo, "A Review on Machine Learning Methods for Customer Churn Prediction and Recommendations for Business Practitioners," in IEEE Access, Vol. 12, pp. 70434-70463, 2024, doi: 10.1109/ACCESS.2024.3402092

[9] Yu Lu, "A Combined Prediction Model of Telecommunications Customer Churn[J]", Journal of Huaqiao University (Natural Science Edition), Vol. 37, No. 5, pp. 637-640, 2016

[10] Cutler, A.; Cutler, D. R.; Stevens, J. R. Random forests. In Ensemble Machine Learning; Springer: Berlin/Heidelberg, Germany, 2012; pp. 157-175

[11] Sperandei, S. (2014) Understanding logistic regression analysis. Biochemia Medica, 24(1), 12-18, https://doi.org/10.11613/BM.2014.003

[12] Lohaj, O.; Pella, Z.; Paralič, J. Data analytics methods for analyzing the impact of factors on early detection of cardiovascular risk. In: 2022 IEEE 20th Jubilee World Symposium on Applied Machine Intelligence and Informatics (SAMI 2022) IEEE, 2022, pp. 000249-000254

[13] Sjarif Nna, Azmi NF, Sarkan HM, Sam SM and Osman MZ. Predicting Churn: How Multilayer Perceptron Method Can Help with Customer Retention in Telecom Industry. IOP Conference Series: Materials Science and Engineering, Volume 864, 2nd Joint Conference on Green Engineering Technology & Applied Computing 2020 4-5 February 2020, Bangkok, Thailand

[14] Liu, Xu-Ying; Wu, Jianxin; Zhou, Zhi-Hua. Exploratory undersampling for class-imbalance learning. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 2008, 39.2: 539-550

[15] Shumaly S.; Neysaryan P. and Guo Y., "Handling Class Imbalance in Customer Churn Prediction in Telecom Sector Using Sampling Techniques, Bagging and Boosting Trees," 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE), Mashhad, Iran, 2020, pp. 082-087, doi: 10.1109/ICCKE50421.2020.9303698

[16] Imani M.; Ghaderpour Z.; Joudaki M. and Beikmohammadi A., "The Impact of SMOTE and ADASYN on Random Forest and Advanced Gradient Boosting Techniques in Telecom Customer Churn Prediction," 2024 10th International Conference on Web Research (ICWR), Tehran, Iran, Islamic Republic of, 2024, pp. 202-209, doi: 10.1109/ICWR61162.2024.10533320

[17] Dobranská L.; Biceková A.; and Babič F., "Classification models comparison from the user's level of interpretability," 2023 IEEE 23rd International Symposium on Computational Intelligence and Informatics (CINTI 2023) Budapest, Hungary, 2023, pp. 000255-000260, doi: 10.1109/CINTI59972.2023.10381999