DOI: https://doi.org/10.15276/aait.08.2025.1
UDC 004.93

Machine learning models for predicting payment status on an online car rental platform

Olena O. Arsirii[1)]
ORCID: https://orcid.org/0000-0001-8130-9613; e.arsiriy@gmail.com. Scopus Author ID: 54419480900
Illia O. Krantovskyi[1)]
ORCID: https://orcid.org/0009-0008-4801-2453; illya20020814@gmail.com
Olexandr V. Rudenko[1)]
ORCID: https://orcid.org/0009-0004-2944-4040; alexo.rudenko@gmail.com
Maria G. Glava[1)]
ORCID: https://orcid.org/0000-0002-9596-9556, glavamg@gmail.com. Scopus Author ID: 57190382998
[1)] Odesa Polytechnic National University. 1, Shevchenko Ave. Odesa, 65044, Ukraine

ABSTRACT

It has been demonstrated that the detailed data collected on online platforms are heterogeneous, semantically inconsistent, and weakly structured. Therefore, the use of machine learning for their aggregation, structuring, and analysis is well-justified. As a case study for developing machine learning models, the task of predicting the payment behavior of clients on an online car rental platform was considered. Input data were automatically generated based on users' actions on the platform. Subsequently, the data were aggregated and structured through feature engineering, time field transformation, and the removal of redundant attributes to enhance model quality. Five classification models were developed: Support Vector Machine, Naive Bayes classifier, Logistic Regression, and two ensemble models (Soft Voting and Stacking). The results showed that Logistic Regression and ensemble models (particularly Stacking) achieved the best precision and recall, making them the most reliable for predicting on-time payments. Ensemble models, especially stacking, demonstrated high efficiency by combining the strengths of different base models. Although SVM can account for complex relationships between features, it showed the weakest performance in distinguishing payment statuses. The findings contribute to a better understanding of customer payment behavior and highlight the importance of choosing appropriate classification models for financial risk assessment. Future research will focus on improving model performance through enhanced feature selection, class imbalance correction, and the integration of additional data sources such as customer credit history. The use of such models can significantly improve automated risk management and enhance decision-making efficiency for companies dealing with payment obligations.

Keywords: Machine learning; payment prediction; naive bayes classifier; logistic regression; support vector machine; ensemble models; financial risk assessment

***For citation*: Arsirii O. O., Krantovskyi I. O., Rudenko O. V., Glava M. G. "Machine learning models for predicting payment status on an online car rental platform". *Applied Aspects of Information Technology*. 2025; Vol.8 No.1: 13–23. DOI: https://doi.org/10.15276/aait.08.2025.1**

INTRODUCTION

Car rental is a popular service that requires effective financial risk management, particularly concerning delayed payments by customers. One of the key aspects of such management is predicting the payment status before the payment request is even sent. This enables rental platforms to take preventive measures to reduce debt levels and improve the company's financial stability. However, the data used for predicting payment status can be automatically collected from various types of user interactions with the platform, including several major sources of information. For instance, this includes customer profile data, rental and payment history, user behavior on the platform, rented vehicles, as well as external factors such as seasonality, economic conditions, and local restrictions. Such data is gathered using web tracking tools (e.g., Google Analytics, Hotjar), which record user behavior on the platform, integrations with payment systems, banking and financial institution APIs (which may grant access to a customer's credit history), the platform's internal database (containing user profile information, booking logs, complaints, reviews), and data from social networks when users sign in via Facebook or Google. It is known that this empirical (raw) data collected from online platforms is typically heterogeneous and weakly structured. To structure and formalize this weakly structured heterogeneous data, various preprocessing methods are applied, such as aggregation, cleaning, filtering, normalization, and encoding [1, 2]. The subsequent application of machine learning models and methods for intelligent analysis of historical customer data opens new possibilities for estimating the probability

of on-time payments or the risk of late payments, ultimately improving the management of accounts receivable.

Recent studies in the field of creditworthiness and payment behavior analysis confirm the effectiveness of machine learning algorithms in automating the prediction of consumers' financial actions [3, 4], [5]. For example, logistic regression and Naive Bayes methods are used in financial and credit institutions to assess clients' creditworthiness. At the same time, the Support Vector Machine (SVM) has proven to be an effective tool for classification and pattern recognition in large datasets [6]. Analyzing factors such as the number of previous overdue payments, the time taken to settle invoices, and the type of service used helps identify patterns that can be leveraged for more accurate payment behavior forecasting.

Additionally, modern approaches to analyzing customer payment behavior involve hybrid models that combine statistical techniques with deep learning. For example, Deep Neural Networks (DNN) and Recurrent Neural Networks (RNN) are increasingly used in fintech for analyzing payment time series [8]. These models are capable of capturing complex dependencies among various parameters, such as service usage frequency, payment history, and behavioral factors, allowing for more accurate risk assessment of late payments and the development of suitable preventive strategies.

LITERATURE OVERVIEW

Predicting customer payment behavior is a critical area of research in financial risk management and the application of machine learning techniques. Advances in data analytics and artificial intelligence have significantly improved the accuracy of payment status prediction, allowing companies to reduce financial risks and optimize debt collection strategies.

In a comprehensive review article, Putrama and Martinek [7] examine key trends in integrating heterogeneous data collected on online platforms across various application domains. Their study focuses on big data rather than a specific research area. It demonstrates that addressing integration challenges related to data semantics and an unstructured format requires the use of advanced technologies such as machine learning.

The role of machine learning in financial risk assessment based on detailed heterogeneous data is thoroughly analyzed. Chang et al. [3] explore the use of artificial intelligence methods to predict payment behavior, emphasizing the importance of feature selection and model interpretability in decision-making. Their study shows that ensemble learning methods, such as boosting, outperform traditional statistical models in financial transaction classification.

Similarly, Lessmann et al. [4] conduct a comparative analysis of machine learning algorithms for credit scoring, showing that support vector machines (SVM) and neural networks provide higher accuracy compared to traditional logistic regression models. Their research highlights the need for robust model validation methods to avoid overfitting and ensure generalizability.

Bayesian methods are also gaining popularity in financial risk management. Senyk et al. [9] propose a Bayesian network (BN)-based model for credit risk assessment, using probabilistic graphical models to analyze borrower behavior. Their study demonstrates the effectiveness of Bayesian networks in structuring financial data, detecting interdependencies, and providing transparent risk assessments. The findings suggest that Bayesian approaches can outperform traditional credit risk models by accounting for uncertainty and variable dependencies, thus improving default prediction accuracy.

Another critical aspect of payment behavior prediction is handling imbalanced datasets, as default cases are often underrepresented. Ozbayoglu et al. [8] present a comprehensive review of deep learning applications in the financial sector, analyzing its effectiveness in areas such as risk assessment and financial decision forecasting. The study categorizes deep learning models–such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), long short-term memory (LSTM), and autoencoders (AEs)–based on their performance in various financial tasks. The results indicate that deep learning models, especially those accounting for temporal dependencies, can significantly enhance classification accuracy and forecasting in imbalanced financial datasets.

Shi et al. [6] conduct a systematic review of credit risk models based on machine learning, outlining the strengths and weaknesses of statistical, classical machine learning, and deep learning approaches. The study identifies issues such as data imbalance, model transparency, and dataset inconsistency as major challenges in credit risk assessment. It concludes that deep learning models, particularly ensemble methods, outperform traditional approaches in credit risk prediction, but also stresses the importance of developing

explainable AI models to improve trust and regulatory compliance.

In recent years, there has been growing interest in using advanced machine learning methods for credit scoring to improve prediction accuracy and interpretability. Xia et al. [10] propose a credit scoring approach based on boosted decision trees, combining gradient boosting (XGBoost) with Bayesian hyperparameter optimization. This model addresses common ensemble method challenges, such as hyperparameter tuning and model interpretability. Results show that the proposed model outperforms traditional methods in terms of accuracy, error rate, and AUC score. Additionally, the use of feature importance and decision plot visualizations improves model transparency, making it more suitable for banking decisions.

Logistic regression remains a widely used method for credit risk assessment due to its interpretability and effectiveness in binary classification tasks. Abid [11] applies logistic regression to evaluate default risk determinants among service-sector companies. The study identifies key financial indicators – such as debt ratio, solvency, profitability, and loan size – as significant predictors of credit risk. Moreover, macroeconomic variables, including inflation rate and GDP growth, play an important role in determining default probability. The findings reaffirm the importance of logistic regression as a baseline tool for credit risk assessment, particularly in regulatory environments where model transparency is a key requirement.

AIMS AND OBJECTIVES OF THE RESEARCH

The aim of the research is to develop machine learning models to enhance the accuracy of payment status prediction based on automatically collected heterogeneous data from customers of an online car rental platform. The implementation of these models will help identify patterns in user payment discipline and improve the company's financial risk management.

To achieve this aim, the following key objectives must be addressed.

1) To form, interpret, and assess the quality of the automatically collected input set of heterogeneous data from online car rental platform customers for further payment status evaluation.

2) To perform data cleaning, normalization, and transformation for subsequent use of machine learning algorithms.

3) To analyze key characteristics that may influence a customer's decision to repay the debt (e.g., the number of previous overdue payments, delay in message delivery, amount of debt, etc.).

4) To apply algorithms (Naive Bayes, Support Vector Machine (SVM), and Logistic Regression) to build corresponding machine learning models for classifying payments as "paid" or "overdue." For final integrated decision-making, to build two ensemble models based on soft voting and stacking.

5) To evaluate the performance of each developed machine learning model using quality metrics (AUC, accuracy, F1-score, specificity, etc.).

6) To present the obtained results in the form of charts and analytical reports, this will allow drawing conclusions about the effectiveness of the approaches used.

The completion of these tasks will not only enable the prediction of the status of future payments but also suggest strategies for reducing overdue debts in the car rental industry.

MAIN PART

Further development of machine learning models was carried out using the input dataset, which contains detailed information about payment requests from customers of the car rental service. The input data were automatically collected by the authors from the online platform [12]. The dataset includes the following input attributes: unique payment request identifier (payment_id), debt amount in cents (amount), currency of the payment (currency), timestamps of request creation (requested_at) and expiration (expires_at), numerical department code (branch), anonymized customer hash code (customer), rental start date (rental_start) and end date (rental_end), payment attempt number (attempt), charge type (product) (e.g., rental fee, damage fee, traffic violation fee), email delivery status (delivery_status) and its corresponding timestamp (delivered_at), and successful payment timestamp (paid_at). The target attribute used is the final payment status (status) (paid or overdue). The dataset size is 19,034 rows.

Data preprocessing was performed using the Pandas library to ensure proper formatting and extraction of useful features.

Before modeling, preliminary data preprocessing was done. All timestamps were converted to the datetime format to ensure correct calculation of time intervals. Data aggregation by unique customer identifier was then performed. During the aggregation process, the paid payment ratio (paid_ratio) was calculated as the average payment status for each customer. Additionally, the average delay between successive payment requests (avg_delay) was calculated, defined as the difference

in days between the request creation timestamps for each customer. These new features were added to the main dataset.

The next step was the development of additional features that could improve model quality. A new variable, rental_duration_days, was created, which reflects the number of days between the start and end of the rental period. Additionally, delivery_delay_hours was calculated, showing the email delivery delay in hours by computing the difference between the timestamps delivered_at and requested_at.

Since some variables in the dataset were categorical, they were encoded into numerical values. The transformation of categories was performed for variables such as product (payment type), branch (company department), and delivery_status (email delivery status). This allowed the proper use of these variables in machine learning models.

An important step before modeling was currency unification. Since the dataset contained payments in different currencies, they were converted into US dollars based on fixed exchange rates: 1 USD = 1.0568 EUR; 1 USD = 1.2065 GBP; 1 USD = 1.0758 CHF. After conversion, all payments were represented in a single currency.

After all transformations, the dataset was cleaned. Columns that no longer contained useful information or duplicated the created features, such as requested_at, expires_at, rental_start, rental_end, paid_at, payment_id, delivered_at, and customer, were removed. This simplified the dataset structure and helped avoid potential multicollinearity in the model.

To ensure that each variable had the same weight, numerical data were standardized. StandardScaler was applied to the columns amount, delivery_delay_hours, and rental_duration_days, normalizing the values of these features and making them more suitable for machine learning models.

The dataset was then split into training and testing sets in an 80 % to 20 % ratio.

Since the original dataset exhibited a significant class imbalance between payment statuses, the Synthetic Minority Over-sampling Technique (SMOTE) was applied. The training dataset showed a significant predominance of payments with the “expired” status (12,711 records) compared to “paid” (2,516 records). The use of SMOTE allowed the synthetic increase of “paid” class records to match the “expired” class, which helped improve the overall performance of the machine learning models.

CHOICE OF DATA ANALYSIS ALGORITHMS

Logistic regression is a widely used supervised machine learning algorithm that belongs to the family of linear regression models but is specifically designed for classification tasks. Unlike linear regression, which predicts continuous values, logistic regression estimates the probability that a given observation belongs to a specific category. In this study, logistic regression is used to classify customers based on their payment behavior, distinguishing those who are likely to make a payment (“reliable customers”) and those who may default on a payment (“unreliable customers”).

The core of logistic regression is the sigmoid function, which transforms any real number into a probability between 0 and 1.

It is expressed as:

$$f(x) = \frac{1}{1 + e^{-x}},$$

where e is the base of the natural logarithm. This transformation ensures that the model outputs a probability, allowing classification based on a predefined threshold (typically 0.5). If the probability exceeds 0.5, the observation is classified as a positive case (e.g., overdue payment), while below 0.5 it is classified as negative (successful payment).

Mathematically, logistic regression is described by the equation [13, 15]:

$$y = \frac{e^{(b_0 + b_1 x)}}{1 + e^{(b_0 + b_1 x)}},$$

where x is the input features; y is the predicted probability; b_0 is the intercept; b_1 is the coefficient for the input feature.

Logistic regression works by adjusting its weight coefficients using iterative optimization methods such as gradient descent to minimize the error between predicted and actual values. The model assigns weights to the input features based on their influence on payment behavior, such as rental duration, number of payment attempts, and transaction history.

Due to its simplicity and interpretability, logistic regression remains a reliable baseline model for binary classification tasks, such as predicting payment status. It is particularly useful in financial applications, where model explainability is crucial for understanding the factors affecting payment probability.

The Naive Bayes classifier is a supervised machine learning algorithm based on Bayesian

statistics, specifically Bayes' theorem. It is designed for classification tasks and operates under the assumption that all features are conditionally independent given the class label, simplifying the computation. Despite this "naïve" assumption of independence, the classifier often performs effectively in various applications [14].

Bayes' theorem is formulated as:

$$P(Y|X) = \frac{P(X|Y) * P(Y)}{P(X)},$$

where $P(Y|X)$ is the posterior probability of class Y given predictor X, $P(X|Y)$ is the likelihood of predictor X given class Y, $P(Y)$ is the prior probability of class Y, and $P(X)$ is the prior probability of predictor X.

In practice, the Naive Bayes classifier computes the posterior probability for each class and assigns the observation to the class with the highest probability. This approach is particularly effective in tasks like text classification, where the model assesses the probability of a document belonging to a specific category based on word frequency [16].

In the context of payment prediction, the Naive Bayes classifier can be used to estimate the probability of payment default based on historical data. Features such as payment history, transaction amount, and customer demographic data can serve as input parameters for the model, enabling businesses to assess credit risk and make informed decisions.

Support Vector Machines (SVM) is a powerful supervised learning algorithm used for classification and regression tasks. Support Vector Machines is particularly effective in high-dimensional spaces and is widely applied in financial analytics, including credit risk assessment, fraud detection, and payment status prediction.

The primary goal of SVM is to find the optimal hyperplane that best separates the data between classes. For two classes, SVM finds a hyperplane that maximizes the distance (margin) between the closest points of each class, known as support vectors. The larger the margin, the better the classifier's ability to generalize [17].

Mathematically, the hyperplane is defined by the equation:

$$w * x + b = 0,$$

where w is the weight vector; x is the input feature vector; b is the bias.

For a given dataset (x_i, y_i), where x_i is the features, and y_i – class labels ($y_i \in \{-1,1\}$), SVM solves the following optimization problem:

$$min_{w,b} \frac{1}{2}||w||^2$$

subject to:

$$y_i(w * x_i + b) \geq 1, \forall i .$$

This ensures that the data is correctly classified with the maximum margin [18].

If the data is not linearly separable, SVM uses kernel functions to transform the data into a higher dimension, where a separating hyperplane can be found.

Popular kernel functions include.

1. Linear kernel

$$K(x, x') = x * x'.$$

2. Polynomial kernel [19]

$$K(x, x') = (x * x' + c)^d.$$

3. RBF kernel (Radial Basis Function) [20]

$$K(x, x') = e^{-\gamma||x-x'||^2}.$$

4. Sigmoid kernel [21]

$$K(x, x') = tanh(\beta x * x' + c).$$

Support Vector Machine is particularly useful for payment status prediction because it can find the optimal boundary between "paid" and "overdue" transactions, effectively handling imbalanced data and complex relationships between features.

Ensemble methods, such as Voting and Stacking, are used to improve prediction accuracy by combining the decisions of several base models. The Voting classifier combines the predictions of different algorithms by applying a voting rule. There are two main approaches: hard voting, where each model makes its choice and the final prediction is determined by the majority vote, and soft voting, where the class probabilities from each model are considered, and the final choice is based on averaged values. This helps smooth individual errors from classifiers and makes the prediction more stable. In the context of payment status prediction, the Voting classifier can combine logistic regression, support vector machines, and Naive Bayes classifiers, ensuring a balance between interpretability, generalization, and robustness to selective anomalies.

The Stacking classifier, unlike Voting, uses a meta-model to combine the output predictions of base models. First, several different algorithms (e.g., logistic regression, SVM, random forest) are trained on the original data, and their predictions are passed to another model, which learns to find the optimal combined solution. The meta-model is often logistic

regression or a neural network, which analyzes the output probabilities from the first-level models and makes the final prediction. The advantage of Stacking is that it can account for different patterns that may be noticeable to one model but not to another. This makes it effective for complex financial tasks such as predicting the risk of overdue payments, where different machine learning methods can complement each other, improving overall accuracy.

After data cleaning and preprocessing, three classification models were chosen for payment status prediction: Support Vector Machine (SVM), which is effective for classification tasks in high-dimensional spaces; Naive Bayes classifier, which is a probabilistic model for predicting categorical outcomes; and Logistic Regression, which is widely used for binary classification tasks.

Parameters for SVM and Logistic Regression were chosen automatically using random search. This approach allows randomly selecting hyperparameter values from a predefined range and evaluating their effectiveness.

For modeling the probability of payment default, the Gaussian Naive Bayes classifier was chosen because it is well-suited for data that follows approximately a normal distribution. Unlike other Naive Bayes variants such as multinomial or Bernoulli, the Gaussian classifier assumes that each feature follows a normal (Gaussian) distribution, which is appropriate for financial data such as transaction amount, average time between payments, and payment frequency. An additional advantage of this algorithm is its resilience to high-dimensional features and low computational requirements, making it efficient for rapid classification.

To improve classification accuracy, ensemble methods were used. The Soft Voting classifier combines the predictions of all models by averaging the probabilities of predicted classes, resulting in a more balanced decision. Stacking was also applied with logistic regression as the meta-model. In this approach, the base models (SVM, Naive Bayes classifier, logistic regression) first make their predictions, after which logistic regression learns to combine these predictions to generate the result.

MODEL EVALUATION

Several metrics were considered for evaluating the performance of the machine learning models used for payment status prediction: Area Under the Curve (AUC), Classification Accuracy (CA), F1-score, Precision, Recall, Matthews Correlation Coefficient (MCC), Specificity (Spec), and LogLoss.

Area under the Curve measures the area under the ROC curve (Receiver Operating Characteristic), which reflects the relationship between sensitivity (Recall) and the rate of false positive predictions (1 - Specificity). A higher AUC value indicates better ability of the model to distinguish between positive and negative classes.

$$\text{AUC} = \int_0^1 TPR(FPR)d(FPR)) \,,$$

where *TPR* (True Positive Rate) is Recall, *FPR* (False Positive Rate) is Specificity.Classification Accuracy (CA) is defined as the ratio of correctly classified samples to the total number of observations in the dataset:

$$CA = \frac{TP + TN}{TP + TN + FP + FN},$$

where *TP* (True Positives) are correctly predicted positive cases; *TN* (True Negatives) are correctly predicted negative cases; *FP* (False Positives) are false positive predictions; *FN* (False Negatives) are false negative predictions.

F1-Score is the harmonic mean between Precision and Recall, providing a balance between the two metrics:

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall}.$$

Precision indicates the proportion of truly positive predictions among all predicted positive cases:

$$Precision = \frac{TP}{TP + FP}.$$

A higher Precision value indicates that the model produces fewer false positive results.

Recall (also known as sensitivity) defines the proportion of correctly predicted positive cases among all actual positive cases in the dataset:

$$Recall = \frac{TP}{TP + FN}.$$

The higher the Recall value, the better the model identifies all positive cases.

Matthews Correlation Coefficient (MCC) considers all four categories of predictions (TP, TN, FP, FN) and provides a more balanced evaluation, especially in cases of imbalanced class distribution:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}.$$

The MCC value ranges from -1 (complete disagreement) to 1 (perfect classification).

Specificity measures the model's ability to correctly classify negative cases:

$$Spec = \frac{TN}{TN + FP}.$$

A higher Specificity indicates fewer false positive predictions from the model.

LogLoss is used to assess the uncertainty in the model's predictions. It is calculated as the average of the logarithmic loss for all predictions:

$$LogLoss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(p_i) + (1 - y_i)\log(1 - p_i)],$$

where N is the total number of samples; y_i is the actual class (0 or 1); p_i is the predicted probability for class 1. The smaller the LogLoss value, the better the model predicts class probabilities.

The following table and graphs summarize the results.

Table. **Quality metrics of developed machine learning models**

Model	AUC	CA	F1	Prec	Recall	MCC	Spec	LogLoss
SVM	0.9980	0.968	0.9042	0.893	0.9157	0.8851	0.9783	0.129
Naive Bayes	0.9977	0.9774	0.9353	0.8873	0.9889	0.9237	0.9751	0.2749
Logistic Regression	0.9984	0.9779	0.9352	0.9085	0.9634	0.9225	0.9808	0.0468
Ensemble (soft voting)	0.9984	0.9764	0.9323	0.8845	0.9857	0.92	0.9745	0.0682
Ensemble (stacking)	0.9983	0.9769	0.9324	0.9019	0.965	0.9193	0.9792	0.0683

Source: **compiled by the authors**

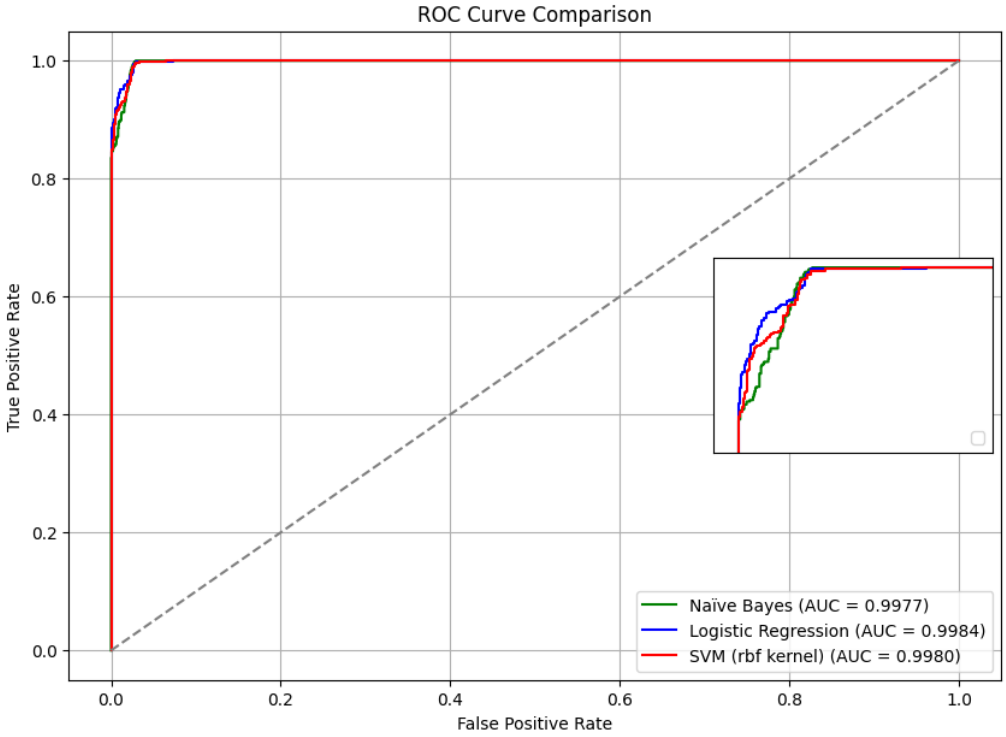

Fig. 1. **ROC curve for Support Vector Machines, Naïve Bayes and Logistic Regression models**
Source: **compiled by the authors**

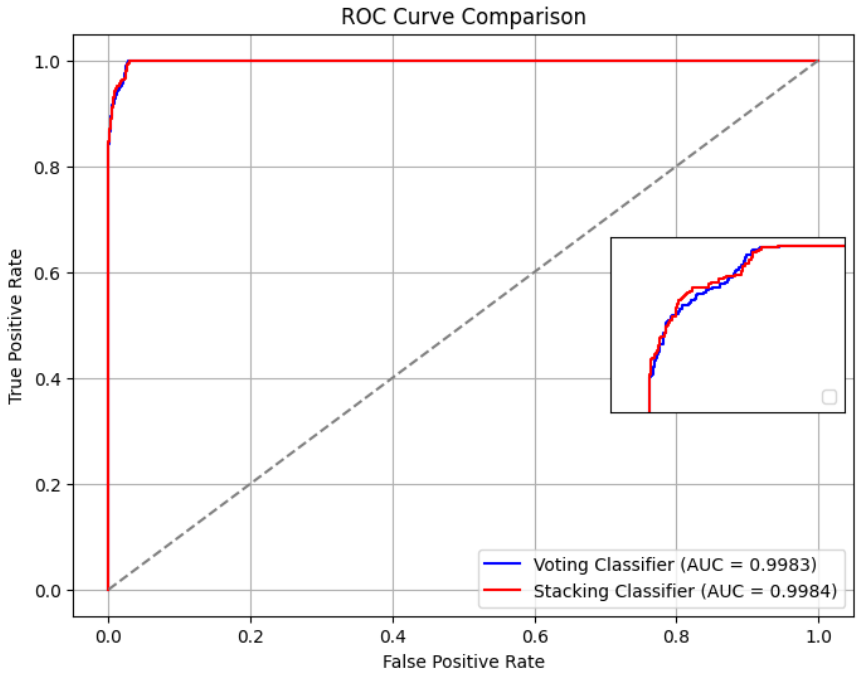

Fig. 2. **Receiver Operating Characteristic curve for ensemble models (soft voting and stacking)**
Source: **compiled by the authors**

According to the results presented, the logistic regression model demonstrates the highest performance among all considered models. It has the highest AUC (0.9984), CA (0.9779), and the lowest LogLoss (0.0468), indicating its high accuracy and reliability. Additionally, it has high values for F1 (0.9352), Precision (0.9085), Recall (0.9634), MCC (0.9225), and Specificity (0.9808), confirming its ability to classify data effectively.

The Naive Bayes and Ensemble (stacking) models also show high results but slightly lag behind logistic regression in some metrics. For example, Naive Bayes has a higher LogLoss (0.2749), and Ensemble (stacking) shows slightly lower Precision (0.9019) and MCC (0.9193) values. The SVM model demonstrates somewhat lower results compared to the other models, especially in terms of F1 (0.9042) and MCC (0.8851).

Based on the results presented, we recommend using the logistic regression model for prediction. This model shows the highest AUC and CA values, indicating its high accuracy and ability to classify data effectively. Additionally, it has the lowest LogLoss, which minimizes classification uncertainty.

Although the Naive Bayes model shows high Recall, which is important for identifying all positive cases, its lower Precision may lead to more false positives. The SVM model, with its lower F1 and MCC values, is not an optimal choice as it does not provide the necessary balance between Precision and Recall.

CONCLUSIONS

In this study, the prediction of payment status (paid or overdue) was analyzed using machine learning models, specifically Support Vector Machines (SVM), Naive Bayes classifier, Logistic Regression, and Ensemble models. The dataset, containing payment requests from car rental service customers, was carefully processed, including feature engineering and removal of redundant attributes to improve model performance.

According to the evaluation results, Logistic Regression demonstrates the highest efficiency in prediction, providing the best balance between Precision and Recall. Its high AUC and CA values and low LogLoss make it a reliable tool for prediction. In contrast to Logistic Regression, the Support Vector Machine (SVM) showed the poorest results in distinguishing payment statuses, highlighting its limited ability to effectively classify financial data. While Naive Bayes and Ensemble models also show high results, Logistic Regression emerges as the optimal choice due to its ability to minimize classification uncertainty and ensure high prediction accuracy. These findings underline the strengths and weaknesses of each model in the context of financial risk assessment and debt management.

REFERENCES

1. Arsirii, O. O., Babilunha, O. Yu., Manikaeva, O. S. & Rudenko, O. I. "Automation of the preparation process weakly-structured multi-dimensional data of sociological surveys in the data mining system". *Herald of Advanced Information Technology. Publ. Science i Technical.* 2018; 1 (1): 11–20. DOI: https://doi.org/10.15276/hait.01.2018.1.

2. Arsirii, O., Antoshchuk, S., Manikaeva, O., Babilunha, O. & Nikolenko, A. "Classification methods of heterogeneous data in intellectual systems of medical and social monitoring". In: *Babichev, S., Lytvynenko, V. (eds) Lecture Notes in Data Engineering, Computational Intelligence, and Decision Making. ISDMCI 2022. Lecture Notes on Data Engineering and Communications Technologies*. *Springer*, Cham. 2023; 149. https://doi.org/10.1007/978-3-031-16203-9_38.

3. Chang V., Sivakulasingam S., Wang H., Wong S. T., Ganatra M. A. & Luo J. "Credit risk prediction using machine learning and deep learning: A study on credit card customers". *Risks*. 2024; 12 (11): 174. DOI: https://doi.org/10.3390/risks12110174.

4. Lessman, S., Baesens, B., Seow, H. & Thomas, L. C. "Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research". *Stochastics and Statistics*. 2015; 247: 124–136, https://www.scopus.com/authid/detail.uri?authorId=6506377858.
DOI: https://doi.org/10.1016/j.ejor.2015.05.030.

5. Arsirii, O. O., Manikaeva, O. S., Nikolenko, A. O. & Babilunha, O. Yu. "Heuristic models and methods for application of the kohonen neural network in the intellectual system of medical- sociological monitoring". *Herald of Advanced Information Technology. Publ. Science i Technical.* 2020; 3 (1): 395–405. DOI: https://doi.org/10.15276/hait.01.2020.6.

6. Shi S., Tse R., Luo W., D'Addona S. & Pau G. "Machine learning-driven credit risk: a systemic review". *Neural Comput & Applic.* 2022; 34: 14327–14339. DOI: https://doi.org/10.1007/s00521-022-07472-2.

7. Putrama, I. M. & Martinek, P. "Heterogeneous data integration: Challenges and opportunities". *Data in Brief*, 2024; 56: 110853, https://www.scopus.com/authid/detail.uri?authorId=57200420805. DOI: https://doi.org/10.1016/j.dib.2024.110853.

8. Ozbayoglu, A. M., Gudelek, M. U. & Sezer, O. B. "Deep learning for financial applications: A survey". *Applied Soft Computing.* 2020; 93,https://www.scopus.com/authid/detail.uri?authorId=57947593100.
DOI: https://doi.org/10.1016/j.asoc.2020.106384.

9. Senyk, A. P., Manziy, O. S., Ohloblin, P. E., Senyk Y. A. & Krasiuk O. P. "Application of the Bayesian approach to modeling credit risks". *Mathematical modeling and Computing.* 2024; 11: 1025–1034. DOI: https://doi.org/10.23939/mmc2024.04.1025.

10. Yufei, X., Chuanzhe, L., YuYing, L. & Nana, L. "A boosted decision tree approach using Bayesian hyper-parameter optimization for credit scoring". *Expert Systems with Applications*. 2017; 78: 225–241, https://www.scopus.com/authid/detail.uri?authorId=57193351080.
DOI: https://doi.org/10.1016/j.eswa.2017.02.017.

11. Abid Lobna. "A logistic regression model for credit risk of companies in the service sector". *International Research in Economics and Finance*. 2022; 6. DOI: https://doi.org/10.20849/iref.v6i2.1179.

12. "Sixt car rental". – Available from: https://www.sixt.com/ – [Accessed: Dec. 2024].

13. Starbuck, C. "The fundamentals of people analytics". 2023. DOI: https://doi.org/10.1007/978-3-031-28674-2.

14. Rish, I. "An empirical study of the naïve bayes classifier". *Empir. methods Artif. Intell.* 2001; 3: 41–46. – Available from: https://www.researchgate.net/publication/228845263_An_Empirical_Study_of_the_Naive_Bayes_Classifier. – [Accessed: Dec. 2024].

15. Peng, C. Y. J., Lee, K. L. & Ingersoll, G. M. "An introduction to logistic regression analysis and reporting". *The Journal of Educational Research*. 2002; 96 (1): 3–14. DOI: https://doi.org/10.1080/00220670209598786.

16. Peretz, O., Koren, M. & Koren, O. "Naive Bayes classifier – an ensemble procedure for recall and precision enrichment". *Engineering Applications of Artificial Intelligence*. 2024; 136 (B), https://www.scopus.com/authid/detail.uri?authorId=57800240100.
DOI: https://doi.org/10.1016/j.engappai.2024.108972.

17. Srivastava, D. & Bhambhu, L. “Data classification using support vector machine”. *Journal of Theoretical and Applied Information Technology.* 2010; 12: 1–7.

18. Cervantes, J., Garcia-Lamont, F., Rodriguez-Mazahua, L. & Lopez, A. “A comprehensive survey on support vector machine classification: Applications, challenges and trends”. *Neurocomputing.* 2020; 408: 189–215, https://www.scopus.com/authid/detail.uri?authorId=23033927200.
DOI: https://doi.org/10.1016/j.neucom.2019.10.118.

19. Vinge, R. & Mckelvey, T. “Understanding support vector machines with polynomial Kernels”. *27th European Signal Processing Conference.* 2019. p. 1–5. DOI: http://dx.doi.org/10.23919/EUSIPCO.2019.8903042.

20. Scholkopf, B., Sung K., Burges, C. J. C., Girosi, F., Niyogi, P. & Poggio, T. “Comparing support vector machines with Gaussian Kernels to radial basis function classifiers”. *Transactions on Signal Processing.* 1997; 45 (11): 2758–2765. DOI: http://dx.doi.org/10.1109/78.650102.

21. Lin, H. & Lin, C. “A study on sigmoid Kernels for SVM and the Training of non-PSD Kernels by SMO-type Methods”. *Neural Computation.* 2003. URL:
https://www.researchgate.net/publication/2478380_A_Study_on_Sigmoid_Kernels_for_SVM_and_the_Training_of_non-PSD_Kernels_by_SMO-type_Methods.

Conflicts of Interest: The authors declare that they have no conflict of interest regarding this study, including financial, personal, authorship or other, which could influence the research and its results presented in this article

Received 10.01.2025
Received after revision 17.03.2025
Accepted 21.03.2025

DOI: https://doi.org/10.15276/aait.08.2025.1
УДК 004.93

Моделі машинного навчання для прогнозування статусу платежів на онлайн-платформі оренди авто

Арсірій Олена Олександрівна[1)]
ORCID: https://orcid.org/0000-0001-8130-9613; e.arsiriy@gmail.com. Scopus Author ID: 54419480900
Крантовський Ілля Олександрович[1)]
ORCID: https://orcid.org/0009-0008-4801-2453; illya20020814@gmail.com
Руденко Олександр Володимирович[1)]
ORCID: https://orcid.org/0009-0004-2944-4040; alexo.rudenko@gmail.com
Глава Марія Геннадіївна[1)]
ORCID: https://orcid.org/0000-0002-9596-9556; glavamg@gmail.com. Scopus Author ID: 57190382998
[1)] Національний університет “Одеська політехніка”, пр. Шевченка, 1. Одеса, 65044, Україна

АННОТАЦІЯ

Показано, що деталізовані дані, які збираються на онлайн платформах є гетерогенними семантично неоднорідними та слабко структурованими. Тому є виправданим використання машинного навчання для їх агрегації, структуризації та аналізу. Як приклад для розробки моделей машинного розглянуто задачу прогнозування платіжної поведінки клієнтів онлайн платформи оренди авто. На основі дій користувачів цієї платформи автоматично формувалися вхідні данні. В подальшому дані агрегувалися та структуризувалися шляхом створення нових ознак, перетворення часових полів та видалення надлишкових атрибутів для підвищення якості моделей. Було розроблено п'ять класифікаційних моделей: метод опорних векторів (support vector machine), наївний байєсівський класифікатор, логістичну регресію та дві ансамблеві моделі (м'яке голосування та стекування). Результати показали, що логістична регресія та ансамблеві моделі (стекування) забезпечили найкращі показники точності та повноти, що робить їх найбільш надійними моделями для прогнозування своєчасних платежів. Ансамблеві моделі, особливо стекування, показали високу ефективність, поєднуючи переваги різних базових моделей. Метод опорних векторів, хоча і здатний враховувати складні взаємозв’язки між ознаками, продемонстрував найгіршу ефективність у розрізненні статусів платежів. Отримані результати дозволяють краще зрозуміти платіжну поведінку клієнтів та підкреслюють важливість правильного вибору класифікаційних моделей для оцінки фінансових ризиків. Подальші дослідження будуть спрямовані на оптимізацію продуктивності моделей шляхом розширеного вибору ознак, усунення дисбалансу класів та інтеграції додаткових джерел даних, таких як кредитна історія

клієнтів. Використання цих моделей може значно покращити автоматизоване управління ризиками та підвищити ефективність прийняття рішень для компаній, що працюють з платіжними зобов'язаннями.

Ключові слова: Машинне навчання; прогнозування платежів; наївний байєсівський класифікатор; логістична регресія; метод опорних векторів; ансамблеві моделі; оцінка фінансових ризиків

ABOUT THE AUTHORS

Olena O. Arsirii – Doctor of Engineering Sciences, Professor, Head of Department of Information Systems, Odessa Polytechnic National University. 1, Shevchenko Ave. Odesa, 65044, Ukraine
ORCID: https://orcid.org/0000-0001-8130-9613; e.arsiriy@gmail.com. Scopus Author ID: 54419480900
***Research field*:** Information technology; artificial intelligence; decision support systems; machine learning; neural networks

Арсірій Олена Олександрівна - доктор технічних наук, професор, завідувач кафедри Інформаційних систем. Національний університет "Одеська політехніка", пр. Шевченка, 1. Одеса, 65044, Україна

Illia O. Krantovskyi - Master. Odesa Polytechnic National University. 1, Shevchenko Ave. Odesa, 65044, Ukraine
ORCID: https://orcid.org/0009-0008-4801-2453; illya20020814@gmail.com
***Research field*:** Information technology; artificial intelligence; machine learning

Крантовський Ілля Олександрович – магістр, Національний університет "Одеська політехніка", пр. Шевченка, 1. Одеса, 65044, Україна

Olexandr V. Rudenko - postgraduate, Odesa Polytechnic National University. 1, Shevchenko Ave. Odesa, 65044, Ukraine
ORCID: https://orcid.org/0009-0004-2944-4040; alexo.rudenko@gmail.com
***Research field*:** Information technology; artificial intelligence; machine learning

Руденко Олександр Володимирович– аспірант, Національний університет "Одеська політехніка", пр. Шевченка, 1. Одеса, 65044, Україна

Maria G. Glava - PhD, Associate Professor, Department of Information Systems, Odesa Polytechnic National University. 1, Shevchenko Ave. Odesa, 65044, Ukraine
ORCID: https://orcid.org/0000-0002-9596-9556; glavamg@gmail.com. Scopus Author ID: 57190382998
***Research field*:** Information technology; databases; database management systems; subject area; artificial intelligence

Глава Марія Геннадіївна - кандидат технічних наук, доцент, доцент кафедри Інформаційних систем. Національний університет "Одеська політехніка", пр. Шевченка, 1. Одеса, 65044, Україна