A Hybrid System for Financial Counselling in Fintech Lending Application

A Hybrid System for Financial Counselling in Fintech Lending Application

David Barrientos[1]
orcid.org/0000-0003-3227-0534

Chantelle Cruz[1]
orcid.org/0000-0003-0595-1852

João Fausto Lorenzato[1]
orcid.org/0000-0002-1150-4904

[1]Escola Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil.
E-mail: djbr@ecomp.poli.br

DOI: 10.25286/repa.v7i3.2462

Como citar este artigo pela NBR 6023/2018: David Barrientos; Chantelle Cruz; João Fausto Lorenzato. A Hybrid System for Financial Counselling in Fintech Lending Application. Revista de Engenharia e Pesquisa Aplicada, Recife, v. 7, n. 3, p.73-81.

ABSTRACT

In a world where high connectivity, portability, and speed have become usual demands in every service, traditional banking has been constantly challenged into evolving and finding better solutions through technology. At this point is where fintechs have gained a huge market portion, setting the pace for the future in banking. In this research project, the case of a specific fintech is considered: Justa, a Brazilian business intended to facilitate common banking services to traders, such as debit and credit transactions, loans, and accounts management. For this project a computational intelligence-based system is developed to attempt to predict accurately the possibility of a client of Justa succeeding at achieving a goal set at the very beginning of the partnership. This, based on their declared characteristics and stored information of past clients.

The system was elaborated from a classification approach, considering the widely known benefits of hybrid systems. Various models were tested using parameter selection in preprocessed data, four of them were then picked to participate in a voting ensemble to make predictions: MLP, KNN, Decision Tree, and Naïve Bayes. Results for the hybrid classification system were by objective metrics: 0.609 accuracy, 0.845 recall, 0.577 precision, and 0.676; which indicate an improvement over the ones obtained in each individual model and align with Justa's interests since reflect a system that is best at predicting true positives. Overall, the proposed system achieved satisfactory results with the given data and its limitations. However, it is considered a successful approach.

KEY-WORDS: FinTech; Lending; Financial Counselling;

1 INTRODUCTION

In latest years, finances and banking have taken their own evolution in technology by introducing concepts of computation, predictive analysis and data mining in their own fields, developing tools and services that have gained great popularity in later years not only in countries with solid economic and technologic background, but also in developing ones **[1]**.

Traditional banking and money transferring services would not succeed in areas where access to physical offices is not common or in the part of population that does not find banking bureaucracy appealing for personal and business finances. Instead, a different kind of company has emerged to offer solutions merging technology and financial services in their business model: FinTechs, which have had a major presence in Brazilian market **[2]** by challenging traditional banking by introducing new models to compete with **[3]**.

The term fintech is described as "an acronym for financial technology, combining bank expertise with modern management science techniques and the computer" (Bettinger, 1972) and often involves technologies as cloud computing, mobile internet and artificial intelligence with financial activities **[4]** **[5]**.

The company subject of this research (Justa) offers potential clients simulations and advise on profits according to estimated incomes in sells, using information the client provides while consulting. After the service is hired and the client starts working with Justa, some time of initial evaluation and adjustment is considered to analyze if the expectations created on simulation have been met as it is usually the case in which a client might alter certain pieces of information (e.g. monthly income) to reach for better deals with Justa and get lower fees per transaction, which would result in economical loses and imbalances. Looking to address this issue, this investigation aims to diminish the incidence of inaccurate estimations and client evaluations; therefore, it becomes necessary the existence of a solution designed to analyze information about potential clients to suggest relationships among Justa and them. The objective of the research is, therefore, to provide a data-driven system able to support decisions based on clients potential to become reliable partners of Justa according to their individual characteristics and patterns contained in information stored by the company about similar clients held in the past.

2 THEORETICAL FOUNDATIONS

2.1 THE MODEL

The proposed approach for this research project depends on the performance of well-known simple classifiers to build a hybrid voting system. Explored classifiers are described in the following paragraphs.

First, the Multi-Layer Perceptron (MLP) utilizes a supervised learning technique called back propagation for training. Its multiple layers and non-linear activation distinguish MLP from a linear perceptron. It can distinguish data that is not linearly separable **[6]**.

The K-Nearest Neighbor algorithm is one of the simplest machine learning techniques. It assumes the similarity between the new data and available cases and put the new case into the category that is most similar to the available categories **[7]**.

Decision tree is one of the most powerful and popular tools for classification and prediction. A Decision tree is a flowchart like tree structure, where each internal node denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (terminal node) holds a class label **[8]**.

A Naive Bayes classifier is a probabilistic machine learning model that's used for classification task. The crux of the classifier is based on the Bayes theorem of a priori probabilities **[9]**.

Also, hybrid systems perform fusion of different models overcoming limitations of traditional approaches based on single classifiers. Combined classifier can outperform the best individual classifier since under certain conditions, this improvement has been proven analytically **[10]**.

2.2 RELATED WORK

Machine learning is a specific subset of Artificial Intelligence that trains machines how to learn. Over the past two decades, the rapid growth in mobile computing systems allowed vast amounts of data gathering and transportation. This, alongside the development of new learning algorithms and theory, enabled computers through Machine Learning to learn pattern identification among vast data sets. Machine Learning is already transforming all kinds of industries, from science and technology to commerce, health care, manufacturing, education, finance, and marketing.

The finance industry has long engaged statistical models and predictive analytics to forecast performance. Machine Learning provides a clear opportunity to advance the transformation of the finance industry playing a significant role in various financial processes such as: robo-advisers **[11]**, loan approvals **[12]**, stock forecasts **[13]**, fraud prevention **[14]**, between others. The inclusion of Artificial Intelligence in the Fintech sector is gaining popularity, with Machine Learning applications in Fintechs predicted to be worth up to $7,305.6 million by 2022 **[15]**.

As assessing a customer credibility is a major challenge, the FinTech sector has been using Machine learning to support the decision process over a potential customer. In 2005 Shin et al. presented a bankruptcy prediction model using support vector machine **[16]**. In 2013 Priyanka and Baby **[17]** proposed a Naive Bayesian algorithm for classifying a customer loan score. In 2015 Sivasree and Sunny **[18]** used a Decision Tree Induction algorithm to find the best attributes and provide reliable loan predictions. Hamid and Ahmed **[19]** proposed in 2016 a supervised classification model based on j48 after comparing the results from a j48 algorithm, Bayes Net, and Naive Bayes. In 2017 Arutjothi and Senthamarai **[20]** proposed a credit scoring system using KNN. In 2018 Panigrahi and Palkar **[21]** used a random forest model determine fraud claims.

3 MATERIALS AND METHODS

The means to achieve the expected results in this research involve efforts in two fronts: specialists in computational techniques who develop the models, and stakeholders who are experienced in the business field and contribute with a final-user perspective.

3.1 STAKEHOLDERS

In this application the stakeholder involved has also gained knowledge on computational intelligence and, therefore, becomes a valuable source of feedback on both ends, which places them in the definition of a promoter.

3.2 DATABASE DESCRIPTION

Required information to develop the proposed model is supplied by the company under anonymization for privacy and security purposes. Then, it is passed to the tech area to be pre-processed in order to transform it and allow proper computational manipulation.

The raw database is formed by examples of 12,659 clients, 11 characteristics in each row and an target variable indicating whether said client achieved or not a promised TPV in a test period with Justa.

3.2.1 Data transformation

In order to obtain data in its most suitable form for processing, some variables were first transformed leading to the features presented in Table 1.

Table 1 – Dictionary for attributes in database

BEFORE	AFTER	COMMENT	TYPE
Order ID	-	Not useful	Categorical
Date of approval	Business Life	Date of approval and Business date of creation were merged into a new variable called Business Life, which represents the number of years that business had when making the order request.	Numerical
Monthly Revenue	Monthly Revenue	Potentially Useful	Numerical
Average Ticket	Average Ticket	Potentially Useful	Numerical
Order Status	-	Not useful	Categorical
MCC	MCC	Potentially Useful	Categorical
Promised TPV	Promised TPV	Potentially Useful	Numerical
Debit Percentage	Debit Percentage	Potentially Useful	Numerical
Credit Percentage	Credit Percentage	Potentially Useful	Numerical
UF	UF	Potentially Useful	Categorical
		Changed to a binary	Categorical output

DOI: 10.25286/repa.v7i3.2462

TPV	Goal	variable that reflects if the client achieved the goal (3 months	

Source: Own author.

With such variables, a feature selection process was executed using Kendall's Rank Correlation as a parameter to reject or not the null hypothesis of variables not having correlation to the output. This coefficient is used specifically in numerical input features considering that the output is categorical.

For categorical inputs, a similar process was executed using chi square test. In each test (Kendall’s and chi square) a significance of 5% was set to reject or not the null hypothesis.

Analysis made for both cases are showcased in Table 2 and Table 3.

Table 2 – Hypothesis test using Kendall’s rank correlation on database for numerical feature selection.

Feature	Kendall’s coeff	Ho	p value	Inference
Business Life	0.036	Rejected	0.00001	Correlated
Monthly revenue	0.018	Rejected	0.03591	Correlated
Average ticket	0.042	Rejected	0.00003	Correlated
Debit percentage	-0.012	Failed to reject	0.158	Uncorrelated
Credit percentage	0.012	Failed to reject	0.158	Uncorrelated
Promised TPV	0.018	Rejected	0.04106	Correlated

Source: Own author.

Table 3 – Hypothesis test using chi square on database for categoric feature selection.

Feature	Statistic ≥ critical value	Ho	p value	Inference
MCC	True	Rejected	~0	Correlated
UF	True	Rejected	~0	Correlated

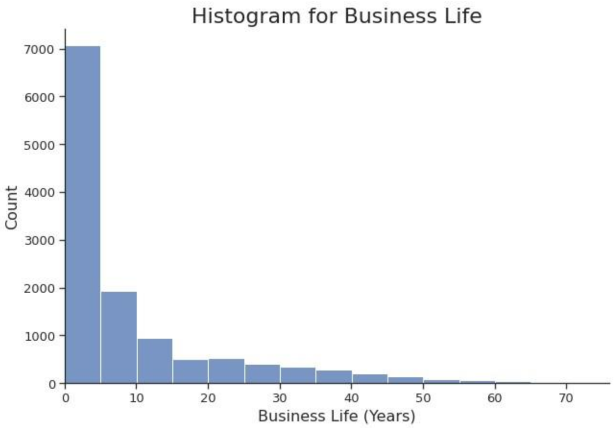
Source: Own author.

3.2 DESCRIPTIVE ANALYSIS

For exploratory purposes on the dataset, some statistical and visual techniques were applied to four numerical variables which have an expected effect on the output (indicating if the client has achieved their set goal or not).

First, graphs were generated indicating the business life of the clients until the date of solicited service. This is presented in Figure 1.

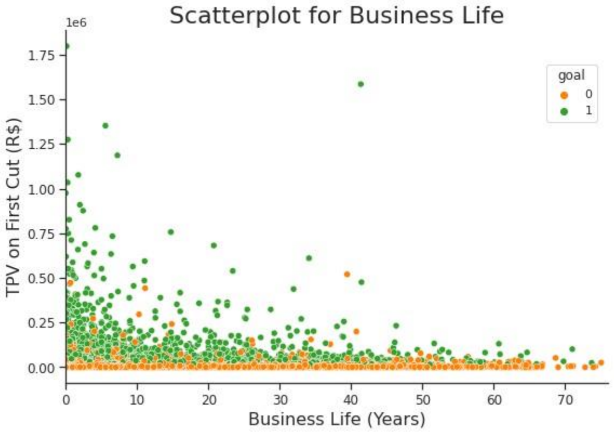
Figure 1 – Client’s business life histogram.



Source: Own author.

Next, graphs were generated for monthly revenue, indicating the average income a client claims to get each month. This could be a major indicator of whether the client is in the capacity of achieving the promised TPV after an observation period. This is presented in Figure 2.

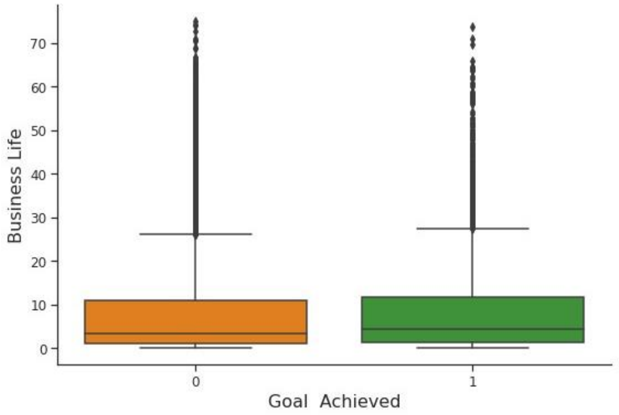
Figure 2 – Client’s business life scatterplot.



Source: Own author.

The next studied variable corresponds to a value also furnished by the client: average ticket indicates the estimated average value a client receives in a transaction, this would suggest the expected TPV per sell and according to the active time of the client, achieving the goal would be expected or not. This is presented in Figure 3.

Figure 3 – Client's business life box plot.



Source: Own author.

3.3 DATA PREPROCESSING

In order to prepare data for its treatment with computational techniques, some preprocessing is needed following five steps:

3.3.1 Cleansing

The database furnished in its initial state includes examples with NaN values which cannot be substituted or inferred from others since it could greatly affect the performance of the system. Observations with a value lower than R$200 for Promised TPV were removed as a R$200 value means that the client would have no goal to achieve, thus, Justa would not sell anything. This value (R$200) was decided alongside Justa.

Outliers were removed by measuring the zScore for each feature. Any z-score greater than 3 or less than -3 is considered to be an outlier. This rule of thumb is based on the empirical rule. From this rule we see that almost all the data (99.7%) should be within three standard deviations from the mean. Shape for the database in this stage was 12590 Observations using 6 Features and 1 Output. Data cleansing steps are summarized in Table 3.

Table 3 – Cleansing process.

ACTION	BEFORE	AFTER	COMMENT
Dropped NANs	12659	12618	41 dropped
Less then $200 TPV	12618	12610	8 dropped
Outliers removal	12610	12590	20 dropped

Source: Own elaboration.

Also, data imbalance was strong in preprocessed base as there was a high number of examples of the class indicating the client did not achieve the goal. This was solved by randomized undersampling, at its final version the database presented a total of 7154 examples.

3.3.2 Coding

Data furnished by Justa follows certain guidelines made to match their recording systems and archives. However, for the purposes of this research, some transformation is needed in variables to allow processing. This involves transforming categorical variables, such as the output, which is presented in values of "Yes" and "No" to determine if the goal set was achieved and is transformed into 1 and 0 values, respectively. Also "UF" and "MCC" suffered transformations via encoders to change label into 5 and 8 bits values, respectively.

3.3.3 Normalization

Since computational techniques perform best with scaled input values a normalization function was applied using the Min-Max principle as noted next as there are no significant outliers to affect the transformation greatly and a range [0.05, 0.95] was chosen to allow the input of smaller or greater values than the ones found on the database. This process was applied to Business life, Average ticket, Monthly revenue and Promised TPV.

3.3.4 Balancing

The database presented two output classes with a high imbalanced (in proportion 1:2.52), a undersampling process was executed. This allowed to reduce the database from 12590 observations to 7146. With values of: 3573 observations for class "0" and 3573 observations for class "1".

DOI: 10.25286/repa.v7i3.2462

3.3.5 Dataset split

The last step in preparation for processing is the division of the dataset into two subsets: training and testing.

The proportion chosen for this system is 85%-15% to allow for the majority of the examples to contribute to training the computational techniques. Training set is later subdivided to obtain a validation set that is meant to help cross validation processes in order to avoid overfitting, and the testing set will serve as a mean to evaluate the performance of the obtained system and its generalization capacity.

3.4 METHODOLOGY

The model developed to solve this classification problem is based on two fundamental statements: there are individual computational methods who perform better than others classifying examples of a database, and the collaboration of the best individual methods promises even better results **[22]**.

Taking these ideas in consideration, individual models were tested to classify the examples in the pre-processed database using KNN, MLP, Decision Tree, and Naïve Bayes classifier.

3.4.1 Metrics

To evaluate the models proposed by this study under different perspectives, three different metrics were selected allowing to make a complete analysis of the final performance of the model since just one function might offer too little information about it. Accuracy, Precision, and Recall, alongside each confusion matrix were measured.

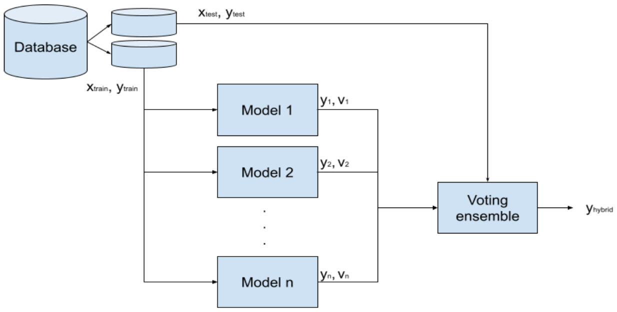
Performances are evaluated using four objective metrics to determine which model suited best the requirements of the study: accuracy, precision, F1, and recall were calculated, along with confusion matrix.

Accuracy is the main metric to consider as it indicates the overall performance of the model predicting correctly. However, recall, precision and F1 help as a support on interpretation of the accuracy.

3.4.2 Hybrid system

After selecting the individual models, a voting ensemble method was implemented to improve overall performance using for each model a combination of accuracy and recall as individual votes and a tiebreaker favoring a "1" value.

Figure 4 – Proposed hybrid system diagram.



Source: Own elaboration.

4 RESULTS AND DISCUSSIONS

4.1 RESULTS

Combinations of parameters were optimized using a grid search with cross validation to avoid overfitting as shown in Table 4. Computational cost for the search of parameters was low as a feature selection was executed in data preprocessing and the size of the database allowed for models to be executed with no obstacles.

Table 4 – Parameters search.

MODEL	SEARCH SPACE	SELECTED PARAMETERS

KNN	**n_neighbors:**
	{5,9,11,13,15}
	weights:
	{uniform, distance}
	metric:
	{Euclidean, Manhattan}
		n_neighbors: 25
		Weights:
		uniform
		Metric:
		euclidean

MLP	**hidden_layer_sizes:** {(50,50,50), (50,100,50), (100,100)}
	activation: {tanh, relu, logistic}
	solver: {sgd, adam, lbfgs}
	alpha: {0.0001, 0.05}
	learning_rate: {constant, adaptive}
		Activation: relu
		Alpha: 0.0001
		Hidden layer sizes: 100, 100
		Learning rate: adaptive
		Solver: Adam

Decision Tree	**criterion:** {gini, entropy}
	Max depth: {4, 5, 6, 7, 8, 9, 10, 11, 12, 15, 20, 30, 40, 50, 70, 90, 120, 150}
		Criterion: gini
		Max depth: 9

SVM	**kernel:** {rbf}
	gamma: {1e-3, 1e-4}
	C: {1, 10, 100, 1000}
		C: 1000
		Gamma: 0.001
		Kernel: RBF

Naïve Bayes	100 numbers spaced evenly on a log scale in range {-9, 0}	**Var smoothing:** 2.848036E-05

Source: Own elaboration based on experiments.

All the individual classifier methods have shown similar performances (Tables 5-8) in the database with accuracies of 0.544, 0.556, 0.528, and 0.514, each for KNN, MLP, DT and Naïve Bayes classifiers, respectively. The metrics for each model are displayed in Table 9.

Table 5 – Confusion Matrix for the KNN Classifier.

	Predicted: 0	**Predicted: 1**
Actual: 0	280	256
Actual: 1	233	303

Source: Own elaboration based on experiments.

Table 6 – Confusion Matrix for the MLP Classifier.

	Predicted: 0	**Predicted: 1**
Actual: 0	308	228
Actual: 1	248	288

Source: Own elaboration based on experiments.

Table 7 – Confusion Matrix for the Decision Tree Classifier.

	Predicted: 0	**Predicted: 1**
Actual: 0	264	272
Actual: 1	234	302

Source: Own elaboration based on experiments.

Table 8 – Confusion Matrix for the Naïve Bayes Classifier.

	Predicted: 0	**Predicted: 1**
Actual: 0	103	433
Actual: 1	88	448

Source: Own elaboration based on experiments.

Table 9 – Evaluation metrics for individual models.

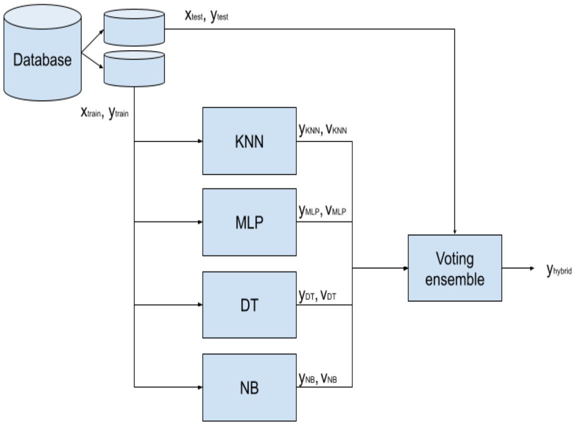
	KNN	**MLP**	**DT**	**Naïve Bayes**
Accuracy	0.544	0.556	0.528	0.514
Recall	0.565	0.537	0.563	0.836
Precision	0.542	0.558	0.526	0.509
F1 score	0.553	0.548	0.544	0.632

Source: Own elaboration based on experiments.

It is to be considered that even when Naïve Bayes has the lowest accuracy, its contribution to the complete model relies on its recall being higher than the rest (0.836, as showcased in Table 15). With this value, the Naïve Bayes model indicates a better "Properly predicted trues" to "Total trues" ratio, which is fundamental for this study given the fact that it is always in Justa's best interest to keep potential good clients, this being a higher priority than avoiding working with an unsuccessful one.

Both accuracy and recall were implemented to determine the vote of each model and fuse them in a voting ensemble.

Figure 5 – Final hybrid system diagram.



Source: Own elaboration based on experiments.

Results from this hybrid system are presented in tables 10 and 11. Here, an improved accuracy of 0.609 is accompanied by a similar recall (0.845) and, therefore, F1 score (0.676) which indicates that along with improvement in accuracy over simple models, the ensemble manages to predict more efficiently in the testing set provided.

Table 10 – Confusion Matrix for the voting ensemble classifier.

	Predicted: 0	**Predicted: 1**
Actual: 0	216	320
Actual: 1	99	437

Source: Own elaboration based on experiments.

Table 11 – Evaluation metrics for voting ensemble classifier.

METRIC	**VOTING ENSEMBLE**
Accuracy	0.609
Recall	0.845
Precision	0.577
F1 score	0.676

Source: Own elaboration based on experiments.

5 CONCLUSIONS AND FUTURE WORKS

The proposed voting ensemble classifier model allows to successfully predict the studied database with accuracy of 0.609, recall of 0.845, precision of 0.577, and F1 score of 0.676. This goes to show an improvement on performance over the individual models (KNN, MLP, Decision Tree, and Naïve Bayes) used to create the final ensemble increasing accuracy level while delivering similar recall and precision ones, ensuring total predictions are better.

Computational cost of the ensemble is also considered similar to the ones of each individual model; therefore, the final model is an acceptable technique to obtain desirable results in prediction of good potential customers for Justa's services and their recommendation.

The presented model leaves opportunity for improvement in different aspects of building process. First, model is likely to get better performance if it is ever possible to train it using features with higher correlation. This would be an aspect depending on the possibility to access other data. Cleansing, balancing, and scaling processes were executed efficiently. However, it is possible to use more complex methods and test if these helps improve final predictions.

For the hybrid system and current conditions, a voting ensemble was the most suitable method to fuse individual models. Other types of ensembles could be considered to improve predictions. Also, the search space parameters for each individual model that build the ensemble can be expanded while considering it will increase computational cost.

REFERENCES

[1] LIEN, N. T. K.; DOAN, T.; and BUI, T. N. **Fintech and banking:Evidence from Vietnam.** The Journal of Asian Finance, Economics, andBusiness, vol. 7, no. 9, pp. 419–426, 2020.

[2] DAROLLES S. et al., **The rise of fintechs and their regulation. FinancialStability Review**, no. 20, pp. 85–92, 2016.

[3] JUNG, D.; DORNER, V.; WEINHARDT, C., & Pusmaz, H. (2018). **Designing a robo-advisor for risk-averse, low-budget consumers**. *Electronic Markets*, *28*(3), 367380.

[4] MILIAN E.Z.; SPINOLA M.; and DE CARVALHO M; **Fintechs: Aliterature review and research agenda**, Electronic Commerce Researchand Applications, vol. 34, p. 100833, 2019.

[5] GODINHO, R.; **Fintech in brazil: Opportunities or threats? InnovativeStrategies for Implementing**

FinTech in Banking. IGI Global, 2021, pp.154–165.

[6] TAUD, H.; MAS, J.F. **Multilayer perceptron (MLP)**. In Geomatic Approaches for Modeling Land Change Scenarios (pp. 451455).

[7] GONGDE, G. et al. **KNN model-based approach in classification**. OTM Confederated International Conferences" On the Move to Meaningful Internet Systems". Springer, Berlin, Heidelberg, 2003.

[8] MYLES, A.J.; FEUDALE, R.N.; LIU, Y.; WOODY N.A.; BROWN, S.D. **An introduction to decision tree modeling**. Journal of Chemometrics: A Journal of the Chemometrics Society, 2004, 18(6), pp.275-285.

[9] RISH, I. **An empirical study of the naive Bayes classifier**. In IJCAI 2001 workshop on empirical methods in artificial intelligence, 2001, Vol. 3, No. 22, pp. 41-46.

[10] MICHAŁ, Woźniak; GRANA, Manuel; and CORCHADO, Emilio. **A survey of multiple classifier systems as hybrid systems**. *Information Fusion* 16 (2014): 3-17.

[11] ARUN, K.; ISHAN, G.; & SANMEET, K. (2016). **Loan approval prediction based on machine learning approach**. *IOSR J. Comput. Eng*, *18*(3), 18-21.

[12] SHEN, S.; JIANG, H.; and ZHANG, T. (2012). **Stock market forecasting using machine learning algorithms**. *Department of Electrical Engineering, Stanford University, Stanford, CA*, 1-5.

[13] SADGALI, I.; SAEL, N.; and BENABBOU, F. (2019). **Performance of machine learning techniques in the detection of financial frauds**. *Procedia computer science*, *148*, 45-54.

[14] **Mediantinc FinTech share**, https://www.mediantinc.com/?aliId=379389, accessed: 2021-10-21.

[15] SHIN, K. S; LEE, T. S.; and KIM, H. J. (2005). **An application of support vector machines in bankruptcy prediction model**. *Expert systems with applications*, *28*(1), 127-135.

[16] PRIYANKA, L. T.; and BABY, N. (2013). **Classification approach based customer prediction analysis for loan preferences of customers**. *International Journal of Computer Applications*, *67*(8).

[17] SIVASREE, M. S.; and SUNNY, T. R. (2015). **Loan credibility prediction system based on decision tree algorithm**. *Int. J. Eng. Res. Technol*, *4*.

[18] HAMID, A. J.; and AHMED, T. M. (2016). **Developing prediction model of loan risk in banks using data mining**. *Machine Learning and Applications: An International Journal (MLAIJ) Vol*, *3*(1)..

[19] ARUTJOTHI, G.; and SENTHAMARAI, C. (2017, December). **Prediction of loan status in commercial bank using machine learning classifier**. In *2017 International Conference on Intelligent Sustainable Systems (ICISS)* (pp. 416-419). IEEE.

[20] PANIGRAHI, S.; and PALKAR, B. (2018). **Comparative analysis on classification algorithms of auto-insurance fraud detection based on feature selection algorithms**. *Int. J. Comput. Sci. Eng*, *6*(9), 72-77.

[21] YUE, Shihong; PING Li; and PEIYI Hao. **SVM classification: Its contents and challenges**. *Applied Mathematics-A Journal of Chinese Universities* 18.3 (2003): 332-342.

DOI: 10.25286/repa.v7i3.2462