International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/IJSRCSEIT

Bank Customer Churn Prediction Using Machine Learning

Dr. Md Jaffar Sadiq[1], Devashish Jobanputra[2], Tadanki Gayithri Sai Kaushik[2], J V V Satya Vrath Rao[2]

[1]Associate Professor, Information Technology Department, Sreenidhi Institute of Science and Technology, Yamnampet, Hyderabad, India

[2]Bachelor of Technology, IT Department, Sreenidhi Institute of Science and Technology, Yamnampet, Hyderabad, India

Article Info
Volume 8, Issue 3
Page Number : 334-341

Publication Issue :
May-June-2022

Article History
Accepted: 01 June 2022
Published:10 June 2022

ABSTRACT

Banking is a very competitive field where customer relations is one of the top priorities for a bank. The bank aims for each customer to be lifelong with them. Home loans are often the bank's longest-term relationship with any customer. According to statistics and some of the top leading banks, customers they require special offers and incentives to retain their engagement with the company. The term 'customer churn' refers to a situation in which a customer or subscriber ceases to transact business with a firm or service provider. To deal with this, many businesses employ machine learning to anticipate the pace at which consumers would churn, and then devise a strategy or offer to keep their current clients.

We use Machine Learning models to forecast customer churn rates, which tells us if a customer is going to stay with the bank or not based on a variety of characteristics. This will assist the bank in determining which customers are most likely to depart. Furthermore, banks can make enticing offers in order to keep their consumers. Well known models such as logistic regression, decision trees, random forest, and various boosting approaches must be utilised in this predictive process to attain a proficient level of accuracy, allowing banks to clearly forecast which customers would depart next based on customer data available.

Keywords : Customer Churn, Machine Learning, Supervised Learning, Gradient Boosting, Banking

I. INTRODUCTION

Customer churn, also known as client attrition, is a word used to describe the likelihood of an established customer continuing to do business with a company. This parameter's probability factor is influenced by a variety of other elements in businesses such as banking, telecommunications, and a few others. In today's harsh market circumstances, it's critical for businesses to measure client churn and the various purposes why customers stop doing business with them.Every business understands that keeping

Copyright: © the author(s), publisher and licensee Technoscience Academy. This is an open-access article distributed under the terms of the Creative Commons Attribution Non-Commercial License, which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited

current customers saves money since acquiring new customers costs five to six times what it costs to keep an existing client. As a result, every company in the market began to study and evaluate the numerous reasons that may lead a consumer or client to abandon the company's services. In fact, in order to keep clients involved with the organisation, the corporation begins to roll out some special presents and incentives for customers who are on the edge of abandoning the company's business.

After numerous big organisationsrecognised the value of customers and customer churn, they began collecting data on customer such as frequency of purchasing a product, and so on. As a result, timely acquisition, storage, and management of client data has become extremely important for businesses. As a result, the deciding process evolved from an event-driven to a data-driven one. Data collection, processing, and analysis were aided by new technologies such as Data Mining, Online storage, and AI. As a result, process was changed to statistical analysis rather than predictive analysis.

II. EXISTINGSYSTEM

In many industries such as telecom, banking, and others, a variety of methodologies have been explored and used to anticipate churn. The majority of the methods made use of data mining techniques and ML algorithms. A large portion of prior research focused on using only one kind of data mining to gather data from appropriate sources, while others focused on a variety of ML methods to assess churn prediction.

Supervised machine learning algorithms have been used in customer churn prediction problems from the days of SVM-POLY using AdaBoost as the best possible model (Vafeiadis, Diamantaras, Chatzisavvas&Sarigiannidis, 2015). The most common algorithms for estimating customer turnover rate include decision tree, random forest, multilayer perceptron, and SVM.

SVM model has been used by the researcher Guo-en, X., for the telecom industry because it can overcome non-linearity, local minimization, and high dimension concerns in available fact-finding of customer churn prediction challenge in practically every feasible field of business. Model prediction is heavily influenced by the data structure and the current situation.

The methodologies that are extensively used to forecast churn turnover are neural networks, SVM, and other regression models. Because neural networks generally outperform traditional statistical techniques such as linear and quadratic discriminant analysis methodologies, data mining research suggests that they should be used for non-parametric datasets.

III. PROPOSED SYSTEM

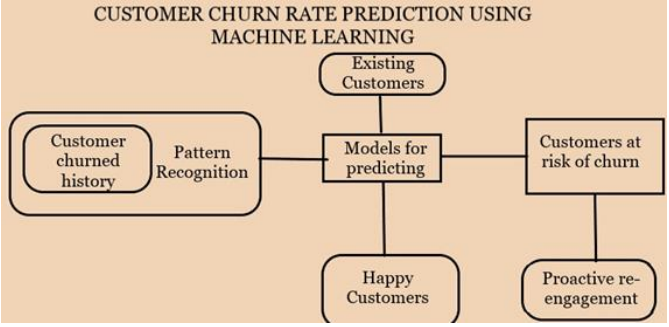
The goal of this project is to employ machine learning and data mining techniques to distinguish between customers who are at risk of being churned and customers who are pleased and satisfied with the bank's products and services. In the following sections, the methods and various technologies used to estimate client turnover for a bank are examined in depth. This project compares each of their prediction power by taking into account the accuracy of each machine learning model and selecting the best model with the highest accuracy, to forecast customer churn. It uses several data pre-processing steps and machine learning algorithms and techniques such as logistic regression, AdaBoost, and others, and compares each of their prediction power by taking into account the accuracy of each machine learning model and selecting the best model with the highest accuracy.

The end-user is given a web app, to enter the data of a customer to determine whether he or she is going to leave or stay with the bank.

ARCHITECTURE

As the volume of the data is continually changing it is an overwhelming task for the analysts to work on

them.At this moment, machine learning-based customer churn prediction enters the picture, and it plays an important role. The prophecy unfolds in stages, each with its own importance. Data is collected from trustworthy sources, pre-processed, and then transformed into a model which can be used to generate predictions. The next phases are to create, test, and eventually implement the model. When compared to prior methods, machine learning algorithms have a high accuracy rate. As indicated in the block diagram, the process is centred on whether the clients are staying with the bank or departing. The procedure of pattern recognition is carried out based on the customers' previous records, and the results are then mapped into the creation of models that are utilised for predicting purposes.



There are two types of customers in the forecast: happy customers and churned customers. Happy customers are those who are satisfied with the bank's fundamental features.

Customers in jeopardy are called by the bank, which attempts to remedy their issues in order to re-engage them, a process known as proactive re-engagement. Customer strategies should be tweaked to increase profitability and encourage more participation.

Proactive support is contingent on the bank's ability to anticipate customer needs before they occur.Pattern recognition is the process of recognising patterns using machine learning algorithms, as shown in the diagram. It can be identified utilising data gathered from the extraction of unique patterns. To assign a label to a pattern, the extracted that is generated using a set of training patterns is employed. As a result, the entire system is built to produce the highest possible accuracy rate, allowing banks to keep customers based on projections.

METHODOLOGY

Thefollowingarethesoftwarerequirementsused in thisprojectandtogettheintended results:

1) Jupyter Notebook
2) Streamlit
3) Visual Studio Code
4) Python Programming

Few of Python libraries which are used to get the results:

a) Pandas
b) Seaborn
c) Matplotlib
d) NumPy

The machine learning models which are used are listed. We calculate the accuracy of themodel's prediction ability and later they are compared and the model with the best accuracy is selected for our prediction purpose.

A. Logistic Regression

A basic categorization technique is logistic regression. It is similar to polynomial and linear regression and belongs to the linear classifiers group. It is great for you to use logistic regression because it is a straightforward and quick strategy for predicting results. Although it is essentially a binary classification method, it may also be applied to multiclass problems like this one.

B. Decision Tree Classifier

A prominent machine learning method is the decision tree. Internal decision-making logic is shared, which isn't the case with black-box algorithms like Neural Networks. It takes less time to train when compared to the neural network approach. The quantity of records and attributes in the data can affect the decision tree's temporal complexity. The decision tree is a non-parametric or distribution-free classifier that doesn't make any assumptions about probability distributions. Decision trees are capable of handling

vast volumes of data while yet producing accurate forecasts.

C. Random Forest Classifier

Random forest is a supervised learning approach. This classifier can be used for regression as well as classification. Random forest is also extremely versatile and straightforward to use. A forest is made up of trees. A forest's strength is thought to increase as the number of trees accessible increases. Random forests build decision trees from randomly picked data samples, get forecasts from each tree, and utilise the voting idea to choose the best possible response. Random forest is also a good way to estimate feature value.

D. K-Neighbors Classifier

K-nearest neighbours (KNN) is a type of KNN algorithm used in supervised machine learning. In its most basic form, KNN is quite simple to design, yet it handles a wide range of classification jobs. Because it lacks a dedicated training phase, it is a sluggish learning algorithm. Rather, while categorising a new data point or instance, it learns on all of the data. Because it makes no assumptions about the raw data, KNN is a nonparametric learning method.

E. AdaBoost Classifier

In recent years, data science and machine learning aficionados have become more familiar with boosting techniques.

Because boosting algorithms are so accurate, some fans want to employ them to win tournaments. The data science projects provide as a platform for learning, investigating, and producing practical solutions to a wide range of business and government issues. Boosting methods combine a number of low-accuracy (or weak) models into a high-accuracy (or strong) model.

F. Gradient Boosting Classifier

Gradient boosting classifiers are a group of machine learning techniques that merge several weak learning models into a single powerful prediction model. Decision trees are widely used in gradient boosting applications. Gradient boosting models are gaining popularity as a result of their efficiency in categorising compound datasets, and have lately been used by data science enthusiasts to win a number of Kaggle competitions. The aim behind "gradient boosting" is to make a sequence of changes to a bad hypothesis or a bad learning algorithm in order to improve the hypothesis's strength.

G. XGBoost Classifier

XGBoost is an open-source software programme that allows you to develop a regularising gradient boosting framework using C++, Java, Python, R, Julia, Perl, and Scala. Linux, Windows, and Mac OS X are all supported. The project's goal is to create a "Scalable, Portable, and Distributed Gradient Boosting (GBM, GBRT, GBDT) Component," according to the website. It may run on a single system as well as distributed processing systems like Apache Hadoop, Spark, and Flink.

STEPS FOR CUSTOMER CHURN PREDICTION

1) Thedatasetwe usehereisacquiredfromKaggleandit goes by“Bankcustomerchurndata”.It has1000records and14features

2) Data pre-processing is the most basic and important step in any machine learning project. Data processing consists of certain stepslike Data cleaning, Data imputation, Dealing with outliers, Data transformation and Data visualization.

3)Thenextstepis to trainthe modeland later the classificationiscaried outusingvariousalgorithms.

4) Validationgivesuswiththeinformation if a customer is going to leave or no.

5) It involved giving values which can give either of the two results that are if the customer is staying or not.

6) If the output is Yes then the customer is retained, and he/she is staying in bank.

7) Ifthe output is Nothen the customerisexitedi.e. churned.

8) Weintendtomakeaweb app weredifferent attributes can be given and it then gives us theoutputwhich can be helpful for the bankwhetherthecustomer/clientis retained or exiting theorganization/bank.

9) This paper intends to predict the best accuracy rate by comparing different algorithms and the output is displayed on the website.

EXPERIMENTALSET UP and RESULTS

There are a number of pre-processing tasks that must be completed before the model can be built effectively. In the final model development and use, the previously mentioned machine learning models are used.

The model construction can be divided into three major parts:

1) Splitting the Predictors(X) and the Target variable(y).

The iloc approach is used in this paper to partition the features into Predictors and Target variables column by column. Inthis case, 10 features are available as X. Those features are:

a) Credit Score
b) Geography
c) Gender
d) Age
e) Tenure
f) Balance
g) Number of products
h) Whether the customer has credit card or not
i) Whether the customer is an active member or not
j) Estimated Salary of the Customer

2)2. Increasing the X and Y split. The test-train split is as follows. We now have four variables. X train and X test; y train and y test The train data represents 80% of the total data, whereas the test data represents 20% of the total data picked at random.

To split the data, we used the train test split python module and set the random state to 42.

3) Using the algorithms, fit the appropriate machine learning model.

Model Training, Testingand Evaluation:

ThealgorithmweuseareimportedfromtheScikitLearnlibrarywhichisanopen-sourcelibrary.

Wetrainedthemodelsusingthetraindatathatwedividedearlierusingthetraintest split.Topredictthevalues,weusethetestdatafromthe traintestsplit.

Forthepurposeofevaluation,weusetheConfusionmatrixandaccuracyparameterderivedfromtheconfusionmatrixparameters.

Accuracyscore=(Number ofCorrectpredictions)÷(TotalNumberofPredictionsmade)

TheevaluationMetricsusedwereTrain/TestsplitvalidationaswellasKfoldCrossvalidationtechniques.

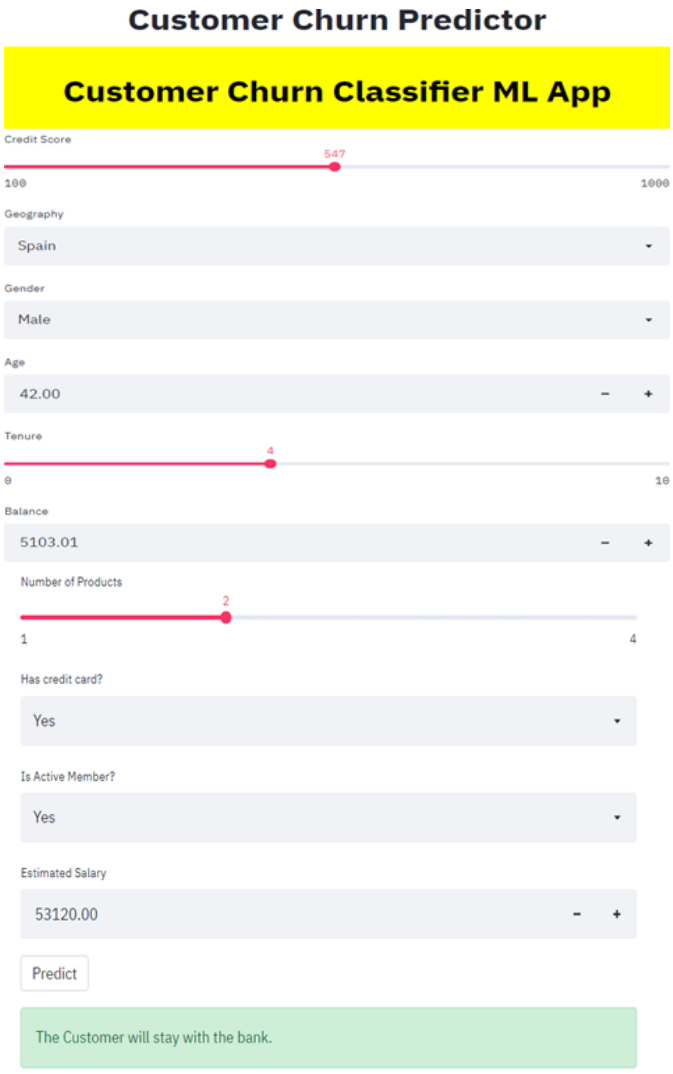
Sincethesetwoaredifferentvalidationtechniques,there willbeaslightdeferenceintheaccuracyscores ofthesamemodel.

With the models used, the accuracies we obtained are:

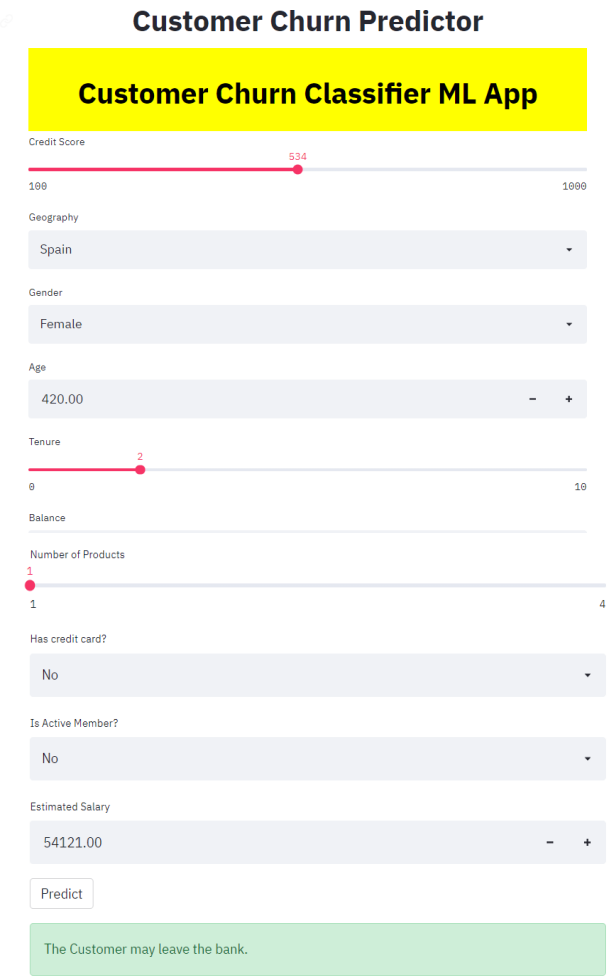
ModelsUsed	Accuracy
LogisticRegression	77.30%
DecisionTreeClassifier	78.93%
RandomForestClassifier	87.22%
K-NNClassifier	83.52%
AdaBoostClassifier	83.55%
GradientBoostingClassifier	84.99%
eXtremeGradientBoostingClassifier	86.72%

Using LOOCV, weobtained the following accuracy scores:

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.8634	0.8593	0.4664	0.7823	0.5840	0.5086	0.5332	0.2260
rf	Random Forest Classifier	0.8587	0.8474	0.4477	0.7711	0.5654	0.4883	0.5146	0.1750
lightgbm	Light Gradient Boosting Machine	0.8578	0.8516	0.4907	0.7313	0.5868	0.5050	0.5199	0.1960
ada	Ada Boost Classifier	0.8573	0.8433	0.4823	0.7352	0.5821	0.5005	0.5169	0.0840
xgboost	Extreme Gradient Boosting	0.8560	0.8390	0.4976	0.7178	0.5874	0.5036	0.5163	0.5200
et	Extra Trees Classifier	0.8403	0.8281	0.4054	0.6929	0.5107	0.4232	0.4452	0.1770
lda	Linear Discriminant Analysis	0.8308	0.8229	0.3098	0.7031	0.4289	0.3472	0.3879	0.0610
ridge	Ridge Classifier	0.8253	0.0000	0.2232	0.7561	0.3438	0.2765	0.3458	0.0410
dt	Decision Tree Classifier	0.7910	0.6898	0.5177	0.4952	0.5053	0.3731	0.3738	0.0140
lr	Logistic Regression	0.7878	0.6667	0.0588	0.3699	0.1006	0.0520	0.0819	0.6880
nb	Naive Bayes	0.7840	0.7543	0.1123	0.4087	0.1757	0.0964	0.1230	0.0080
knn	K Neighbors Classifier	0.7573	0.5218	0.0790	0.2357	0.1177	0.0164	0.0198	0.0300
svm	SVM - Linear Kernel	0.6914	0.0000	0.2202	0.1173	0.1334	0.0235	0.0301	0.0300
qda	Quadratic Discriminant Analysis	0.2062	0.5000	1.0000	0.2062	0.3419	0.0000	0.0000	0.0790



OUTPUT-1



Output-2

IV. CONCLUSION AND FUTURE ENHANCEMENTS

After a long period of low interest rates, the banking sector around the world has been undergoing significant structural changes in its business model in order to reach profitability targets. In order to combat low financial margins, commissions have been significantly increased in order to boost income. There has been a shift in operations and the closure of different branches across the country as a result of the decision to migrate to digital transactions. Certain actions, such as these, have had a substantial impact on customer consideration and turnover rates.

Finally, the strategy suggested in the publication achieved its major goal of accurately and reliably predicting customer attrition rate.

It has offered a mechanism to keep track of consumers and their interactions with the bank in real time. When the bank recognises customers who are at risk of being churned, it can use various schemes and marketing methods to keep the consumer connected and engaged with the bank. As a result, a bank can be successful in maintaining its customers, who will eventually refer new customers to the bank.

The current research reported in this study considers datasets from a specific time period rather than the whole data set since the bank's inception. As a result, the produced findings may be skewed based on the data available at the time of prediction. Finally, the proposed prediction system can be improved by gaining access to diverse datasets from various years since the bank's establishment. As a result, the accuracy and reliability of our machine learning model's prediction has improved.

V. REFERENCES

[1]. Y. Lecun, Y. Bengio, and G. Hinton, "Deep learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015.

[2]. A. Krizhevsky, "ImageNet Classification with Deep Convolutional Neural Networks," NIPS, vol. 4, no. 4, pp. 253–262, 2012.

[3]. A. Nurhadiyatna, S. Cahyadi, F. Damatraseta, and Y. Rianto, "Adult content classification through deep convolution neural network," Proc. - 2017 Int. Conf. Comput. Control. Informatics its Appl. Emerg. Trends Comput. Sci. Eng. IC3INA 2017, vol. 2018-Janua, pp. 106–110, 2018.

[4]. X. Jin, Y. Wang, and X. Tan, "Pornographic Image Recognition via Weighted Multiple Instance Learning," IEEE Trans. Cybern., vol. PP, pp. 1–9, 2018.

[5]. F. Nian, T. Li, Y. Wang, M. Xu, and J. Wu, "Pornographic image detection utilizing deep convolutional neural networks," Neurocomputing, vol. 210, pp. 283–293, 2016.

[6]. K. Zhou, L. Zhuo, Z. Geng, J. Zhang, and X. G. Li, "Convolutional neural networks based pornographic image classification," Proc. - 2016 IEEE 2nd Int. Conf. Multimed. Big Data, BigMM 2016, pp. 206–209, 2016.

[7]. D. Moreira et al., "Pornography classification: The hidden clues in video space–time," Forensic Sci. Int., vol. 268, pp. 46–61, 2016.

[8]. M. D. More, D. M. Souza, and R. C. Barros, "Seamless Explicit images Censorship : an Image-to-Image Translation Approach based on Adversarial Training," IEEE Int. Jt. Conf. Neural Networks, 2018.

[9]. A. P. B. Lopes, S. E. F. De Avila, A. N. A. Peixoto, R. S. Oliveira, M. D. M. Coelho, and A. D. A. Araújo, "Nude detection in video using bag-of-visual-features," Proc. SIBGRAPI 2009 - 22nd Brazilian Symp. Comput. Graph. Image Process., pp. 224–231, 2009.

[10]. R. Shen, F. Zou, J. Song, K. Yan, and K. Zhou, "EFUI: An ensemble framework using uncertain inference for pornographic image recognition," Neurocomputing, vol. 322, pp. 166–176, 2018.

[11]. A. P. B. Lopes, S. E. F. De Avila, A. N. A. Peixoto, R. S. Oliveira, and A. De A. Araújo, "A bag-of-features approach based on Hue-SIFT descriptor for nude detection," Eur. Signal Process. Conf., no. Eusipco, pp. 1552–1556, 2009.

[12]. W. Zhou, A. Ahrary, and S. I. Kamata, "Image description with local patterns: An application to face recognition," IEICE Trans. Inf. Syst., vol. E95-D, no. 5, pp. 1494–1505, 2012.

[13]. D. G. Lowe, "Object recognition from local scale-invariant features," Proc. Seventh IEEE Int. Conf. Comput. Vis., pp. 1150–1157 vol.2, 1999.

[14]. N. Dalal, B. Triggs, and D. Europe, "Histograms of Oriented Gradients for Human Detection," 2005.

[15]. L. Lv, C. Zhao, H. Lv, J. Shang, Y. Yang, and J. Wang, "Pornographic images detection using high-level semantic features," Proc. - 2011 7th Int. Conf. Nat. Comput. ICNC 2011, vol. 2, pp. 1015–1018, 2011.

[16]. J. Zhang, L. Sui, L. Zhuo, Z. Li, and Y. Yang, "An approach of bag-of-words based on visual attention model for pornographic images recognition in compressed domain," Neurocomputing, vol. 110, no. July 2012, pp. 145–152, 2013.

[17]. C. Caetano, S. Avila, S. Guimar, and A. D. A. Ara, "Pornography Detection using BOSSANOVA Video Descripor," pp. 2–6, 2014.

[18]. S. Avila, N. Thome, M. Cord, E. Valle, and A. De A. Araújo, "Pooling in image representation: The visual codeword point of view," Comput. Vis. Image Underst., vol. 117, no. 5, pp. 453–465, 2013.

[19]. C. Caetano, S. Avila, W. R. Schwartz, S. J. F. Guimarães, and A. de A. Araújo, "A mid-level video representation based on binary descriptors: A case study for pornography detection," Neurocomputing, vol. 213, pp. 102–114, 2016.

[20]. D. Li, N. Li, J. Wang, and T. Zhu, "Pornographic images recognition based on spatial pyramid partition and multi-instance ensemble learning," Knowledge-Based Syst., vol. 84, pp. 214–223, 2015.

[21]. C. X. Ries and R. Lienhart, "A survey on visual adult image recognition," Multimed. Tools Appl., vol. 69, no. 3, pp. 661–688, 2014.

[22]. Z. Geng, L. Zhuo, J. Zhang, and X. Li, "A comparative study of local feature extraction algorithms for Web pornographic image recognition," Proc. 2015 IEEE Int. Conf. Prog. Informatics Comput. PIC 2015, pp. 87–92, 2016.

[23]. R. Nejad, Elaheh Mahraban and Affendey, Lilly Suriani and Latip, Rohaya Binti and Ishak, Iskandar Bin and Banaeeyan, "Transferred Semantic Scores for Scalable Retrieval of Histopathological Breast Cancer Images," pp. 1–8, 2018.

[24]. R. Banaeeyan, H. Lye, M. F. Ahmad Fauzi, H. Abdul Karim, and J. See, "Semantic facial scores and compact deep transferred descriptors for scalable face image retrieval," Neurocomputing, 2018.

[25]. K. Fernandes, J. S. Cardoso, and B. S. Astrup, "A deep learning approach for the forensic evaluation of sexual assault," Pattern Anal. Appl., vol. 21, no. 3, pp. 629–640, 2018.

[26]. R. G. Crane, "Deep Residual Learning for Image Recognition 2015," no. (ed.), Oxford, U.K., Pergamon Press PLC, 1989, Section 3, pp.111-120. (ISBN 0-08-036148-X), pp. 1–9, 1989.

Cite this article as :

Dr. Md Jaffar Sadiq, Devashish Jobanputra, Tadanki Gayithri Sai Kaushik, J V V Satya Vrath Rao, "Bank Customer Churn Prediction Using Machine Learning", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 3, pp. 334-341, May-June 2022.
Journal URL : https://ijsrcseit.com/CSEIT228389