International Journal of Advances in Engineering and Management (IJAEM)
Volume 3, Issue 12 Dec 2021, pp: 1066-1072 www.ijaem.net ISSN: 2395-5252

Predicting Customer Churn in a Credit card company by applying ML and AutoML tools.

Amrita Doshi
Student, Shanti Business School Ahmedabad.

Submitted: 05-12-2021 Revised: 17-12-2021 Accepted: 20-12-2021

ABSTRACT:
Every year, many customers buy credit cards, and many customers leave so we need to forecast future churn in the credit card company. Credit card churn prediction can assist a bank to know which of his customer is going to retain and can also assist the bank to know on which customer he needs to emphasize more so that he can earn the highest gains. Future churn can be estimated by evaluating recent industry patterns and churn rate, as well as potential innovations. Total transaction, Total transaction count, and total relationship count matters a lot in retaining the customer. This paper's functionality is based on a website that recognizes customer requirements and then incorporates the use of data mining's multiple regression algorithm and Auto-ML. This study aims to use classification analysis and Auto-ML to predict Customer Churn based.

Keywords: Credit card churn, AutoML, machine learning

I. INTRODUCTION:

In the current business climate, banks and monetary organizations have countless customers. Banks offer types of assistance through different channels, similar to ATMs, Credit cards, Debit cards, web banking, and so forth The quantity of customers has expanded immensely and customers have gotten progressively aware of the nature of administration. This energizes colossal rivalry among different banks, bringing about a huge expansion in the dependability of and nature of administration from banks. Additionally, customers move loyalties starting with one bank then onto the next on account of different reasons, for example, the accessibility of the most recent innovation, customer-accommodating bank staff, low financing costs, the nearness of the geological area, the different administrations offered, and so forth. Subsequently, there is a squeezing need to build up a model that can anticipate which customer is probably going to produce dependent on the customers' segment, psychographic, and value-based information.

Bolton (1998) proposed that administration associations ought to be proactive and gain information from customers before they are imperfect by understanding their present fulfillment levels. He additionally recommended that administration experiences go about an early marker of whether an association's relationship with a customer is prospering or not. He inferred that the customers who have long relationships with the firm have higher earlier combined fulfillment evaluations and more modest resulting in apparent misfortunes that are related to ensuing help experiences. He also proposed that it is hypothetically more productive to section and target customers based on their (changing) buy practices and administration encounters, instead of based on their (steady) socioeconomics or different factors.

II. LITERATURE REVIEW

Rajamohamed and Manokaran (2008) advanced a credit-card churn prediction model built on supervised learning techniques and rough clustering. The customers retention technique for credit card churn prediction (C3P) was completed using supervised classification techniques using K-means algorithm. Lalwani et. al. (2021) proposed a methodology consisting of six phases. Initially, datapre-processing as well as feature analysis were done. The data was split into test and train data in the proportion of 80% and 20% respectively. Logistic regression, Naïve Bayes, Random forest, Support vector machine, decision tree algorithms were applied on the dataset. Also, boosting and ensemble techniques were used to check the effect on the accuracy of the models. The highest accuracy was obtained through Adaboost and XGboost classifiers. Tsiakmari (2020) implemented AutoML for prediction using an educational dataset. The study explored the use of advanced ML techniques by utilizing hyperparameter optimization in an educational setting. The analysis was restricted to tree-based and rule-based models. The study demonstrated that autoML tools consistently outperform other machine learning methods. He, Zhao and Chu (2019) compared various autoML tools based on the pipeline, feature engineering, data preparation, neural architecture search and hyperparameter optimization. CIFAR-10 dataset and ImageNet datasets were used to compare the performance of various NAS algorithms. The problems existing with autoML and future research were also discussed. Truong et al. (2019) compared recent state of autoML tools to automate tasks such as data pre-processing, model selection, feature engineering prediction result analysis and hyperparameter optimization. Different datasets were used in order to evaluate their performance and find their pros and cons. Nie et al. (2011) used logistic regression and decision tree to forecast credit card churn using credit card data from a Chinese bank. The study examined the relevance of variables including card information, customer information, transaction activity and risk information. The study also developed a misclassification cost measurement considering the two types of errors as well as economic sense. The study showed that regression has better accuracy as compared to decision tree.

III. RESEARCH METHODOLOGY

The study uses a secondary data based quantitative method in order to develop a model to predict customer churn. Following are the research objectives:

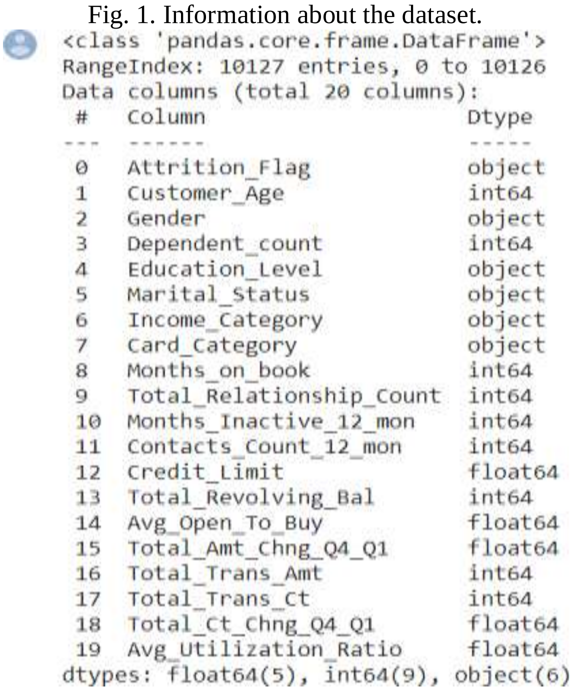
1. Createaneffectivecustomerchurnpredictionmodel.
2. Trytoprovidesomeinterpretability.
3. Toidentifyandvisualizewhichfactors contributetocustomerchurn.
4. Tocreateaneffectivecustomer churnpredictionmodel.
5. Tovalidatethemodel's predictionaccuracy.

The dataset used in the study is Credit card customer dataset extracted from Kaggle website. Following are the variables considered: CLIENTNUM,Attrition_Flag,Customer_Age, Gender, Numberof dependents, Education_Level, Marital_Status,Income_Category, Card_Category, Months_on_book, Total_Relationship_Count,Months_Inactive, Contacts_Count, Credit_Limit,,Total_Revolving_Bal, Avg_Open_To_Buy, Total_Amt_Chng, Total_Trans_Amt, Total_Trans_Ct, Total_Ct_Chng, Avg_Utilization_Ratio.

IV. ANALYSIS

Basic Information about the dataset is provided in fig. 1.

Fig. 1. Information about the dataset.

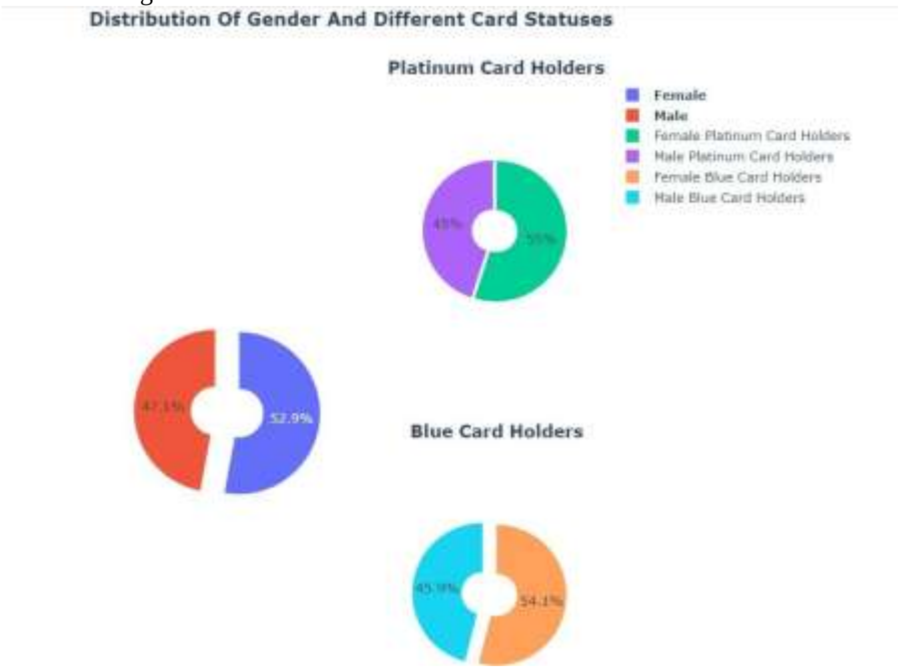


```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 10127 entries, 0 to 10126
Data columns (total 20 columns):
 #   Column                    Dtype
---  ------                    -----
 0   Attrition_Flag            object
 1   Customer_Age              int64
 2   Gender                    object
 3   Dependent_count           int64
 4   Education_Level           object
 5   Marital_Status            object
 6   Income_Category           object
 7   Card_Category             object
 8   Months_on_book            int64
 9   Total_Relationship_Count  int64
 10  Months_Inactive_12_mon    int64
 11  Contacts_Count_12_mon     int64
 12  Credit_Limit              float64
 13  Total_Revolving_Bal       int64
 14  Avg_Open_To_Buy           float64
 15  Total_Amt_Chng_Q4_Q1      float64
 16  Total_Trans_Amt           int64
 17  Total_Trans_Ct            int64
 18  Total_Ct_Chng_Q4_Q1       float64
 19  Avg_Utilization_Ratio     float64
dtypes: float64(5), int64(9), object(6)
```

The data has has a total of 10127 rows and 21 columns. There are three types of data types: integer, float, and object. Generally, the object is the categorical type. Here we do not have any sort of missing values in the data so we do not need to do any pre- processing of the data in perspective of the missing data. The data is modulated as per the requirement of the Automl tools. Here data modulation is done depending upon the columns of the raw data. The description attribute has various information as the Customer has which card, Gender of the customer, Education level of the customer, Income category of the customer, Marital Status of the customer, Attrition Flag. So, for all these columns I have applied 1 and 0; 1 if the customer has and 0 if the customer doesn't have. For example, the id 7039061606 has aPlatinum Card in the description attribute, it put 1 and he isnot married so in the married column it will have 0. Like this method, a total of 24 more columns are created.

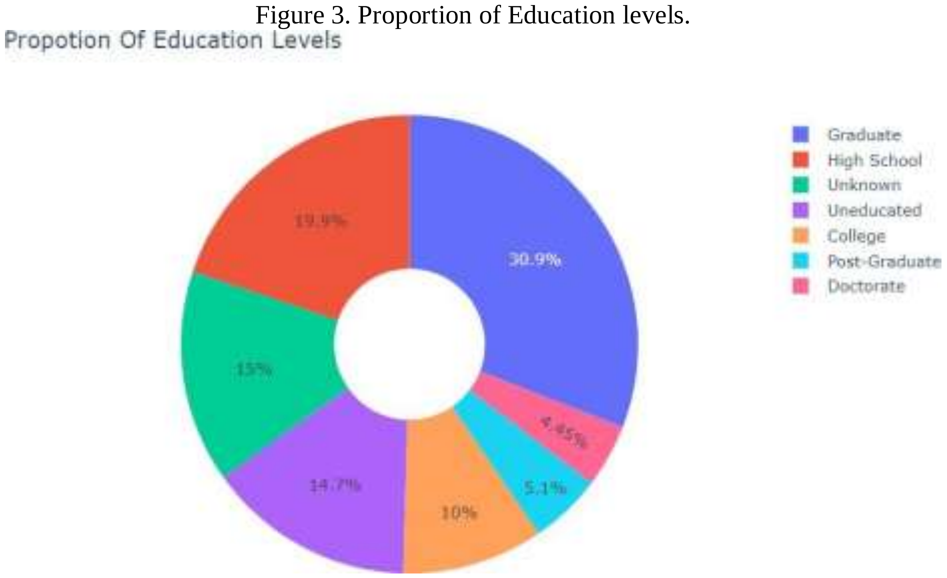
1.1 Exploratory data analysis of credit card churn dataset

Fig. 2. Distribution of Gender and different credit card statuses



From fig. 2, we can see that moresamples offemalesinourdatasetarecomparedtomales,butthepercentageofdifferenceis notthatsignificant,sowecansaythatgenders areuniformlydistributed.

Figure 3. Proportion of Education levels.



From figure 3, we can see that if mostofthecustomers withunknowneducationstatus lackanyeducation,wecanstatethatmore than70% ofthecustomershaveaformaleducationlevel. About35%haveahigherlevelof education.□

Figure 4. Distribution of total number of products held by the customer.

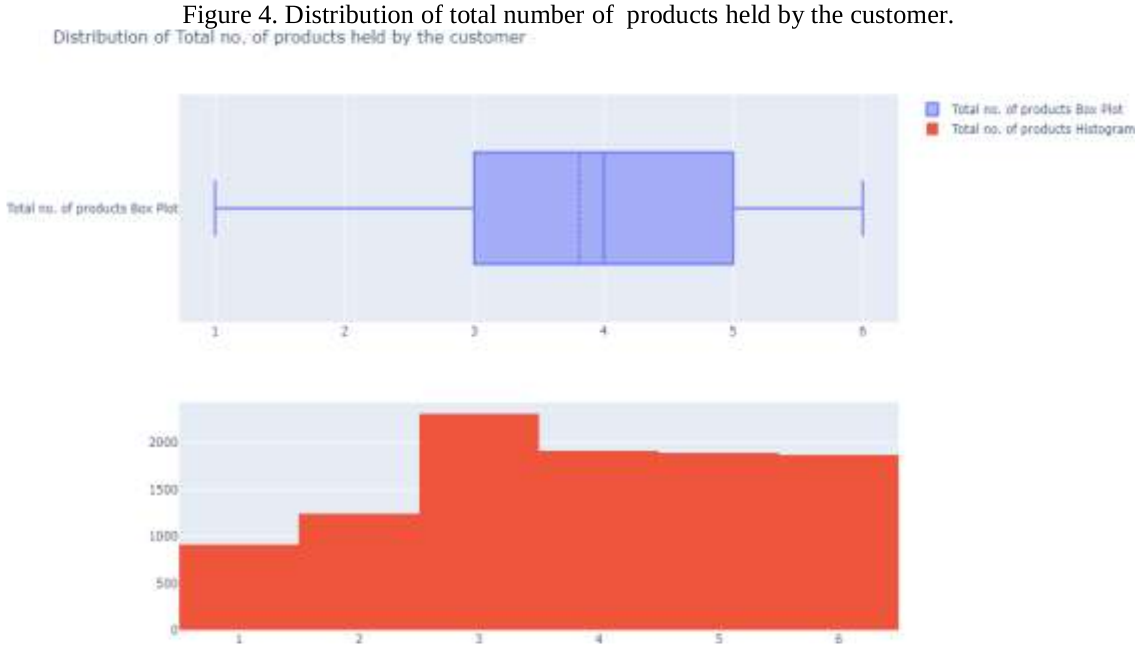
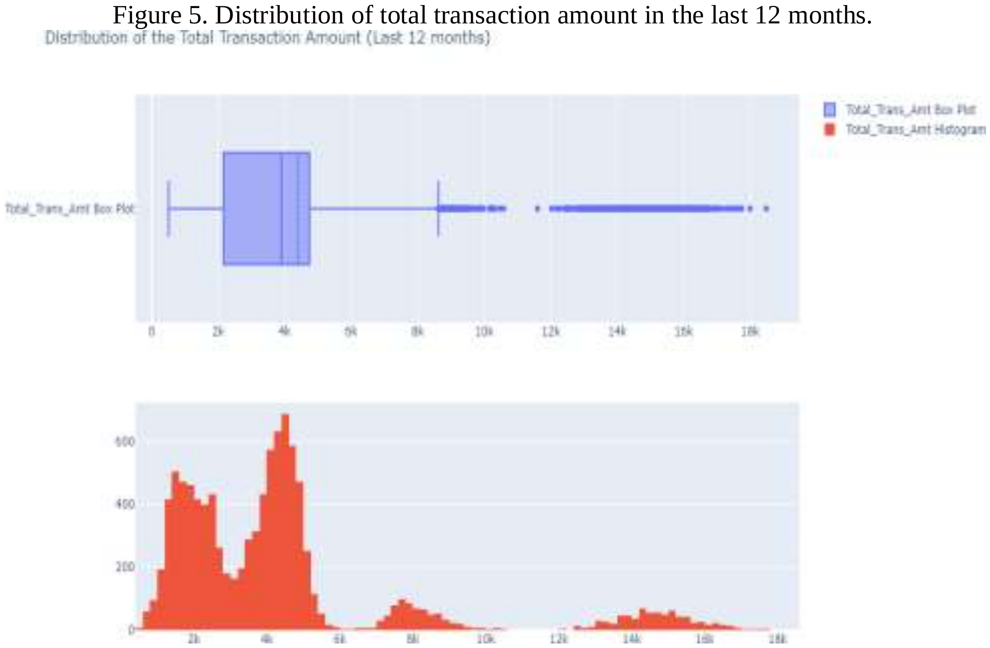


Figure 4 shows that thedistributionofthetotalnumberofproducts heldbythecustomerseems closertoauniform distributionandmayappearuseless as apredictorforchurnstatus.□

Figure 5. Distribution of total transaction amount in the last 12 months.



From figure 5, we can see that the distribution of the total transactions (Last 12 months) displays a multimodal distribution, meaning we have some underlying groups in our data; it can be an interesting experiment to try and cluster the different groups and view the similarities between them and what describes best the different groups which create the different modes in our distribution.

Figure 6. Distribution of the number of months inactive in the last 12 months.

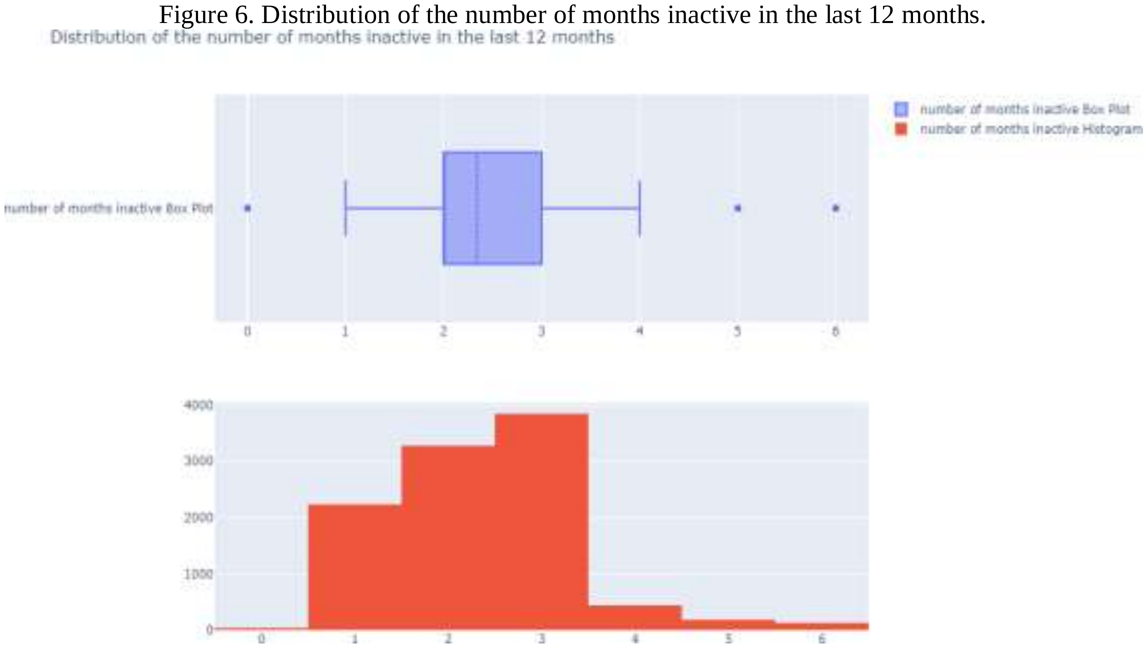
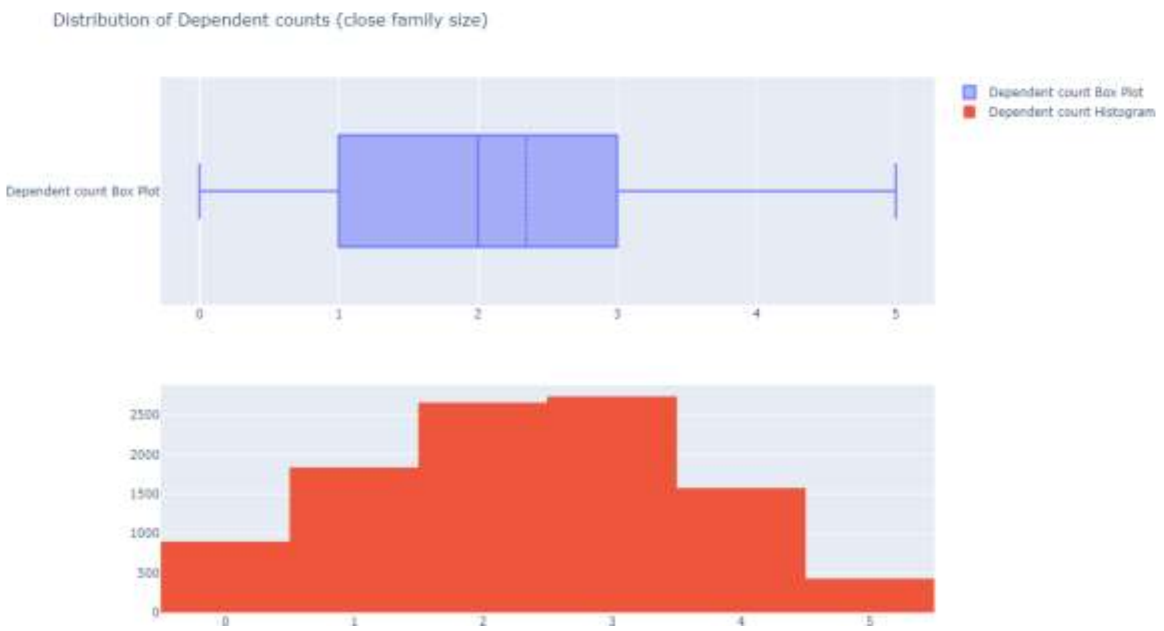


Figure 7. Distribution of dependent counts (close family size)



From fig. 6 and 7, we can see that thedistributionofDependentcountsisfairlynormallydistributedwithaslightrightskew.□

Analysis

Here, four different methods of AutoML were evaluated based on Accuracy values.The accuracy values are provided in table.

Table 1. Accuracy of the AutoML models.

Method	Accuracy(%)
H2O-GradientBoosting	82
H2O-RandomForest	41
H2O-DeepLearning	61.09
Auto-Sklearn	95.9
Auto-Keras	81.11

From table 1, we can see that Auto-Sklearn has the highest accuracy and H20- deep learning has the lowest accuracy.

Managerial Implications

The research finding has important application for companies that are active in the banking industry regarding building a predictive model for Debit Card's customers' churn. Besides the idea of developing the dual-step model for extracting the churn definition before the model building phase, can also be applied in the mobile telecommunications market (Especially the pre-paid ones).

Dealing with a lot of information delivered by clients in organizations and associations can give them valuable information in regards to their clients which can be misused in growing new items, directing maintenance crusades, and strategically pitching and up-selling the items and administrations of an organization. This shows the importance of the utilization of information min ing in showcasing. Indeed mining, the crude information created by clients in their touchpoints with the organization can give the organization superior knowledge toward their clients which encourages them to lead more productive and more compelling advertising ventures.

The goal of this exploration was to build up a prescient model for client stir in the Debit Card's clients' beat which can recognize clients who are probably going to agitate in close fates and the ones who are left with the organization. The commitment of a particular model for the organization is that it would forestall the misuse of cash because of the mass promoting approaches and it empowers the organizations to focus on the genuine churners by removing the clients with a high likelihood of agitating. Plus, as talked about in the introduction the expense of procuring another client is multiple times more than holding a current one, subsequently since the churn prescient model is fit for demonstrating the future churners, the organizations that are expected to keep up their client base can zero in on maintenance approaches rather than obtaining approaches which are less exorbitant.

Summarizing, for the finding of this exploration, we can recommend the organizations to use the data mining strategies to change the current client data in their databases to exploitable information that can help them in their showcasing plans. Besides, it would VIII be valuable for them to construct a prescient churn model by the utilization of data mining which assumes the part of a cautioning framework for the organizations, and furthermore, it can assist them with spending their maintenance financial planproficiently.

Limitations of the study

The period of study is only 2 months. There were limited resource available. Variables are limited as only the variables provided in the website were considered. The persona of the customers were not available as well as alack of customer information. No data on location or geography of the dataset.

REFERENCES:

[1] Rajamohamed, R., &Manokaran, J. (2018). Improved credit card churn prediction based on rough clustering and supervised learning techniques. Cluster Computing, 21(1), 65-77.

[2] Lalwani, P., Mishra, M. K., Chadha, J. S., & Sethi, P. (2021). Customer churn prediction system: a machine learning approach. Computing, 1-24.

[3] Tsiakmaki, M., Kostopoulos, G., Kotsiantis, S., &Ragos, O. (2020). Implementing AutoML in educational data mining for prediction tasks. Applied Sciences, 10(1), 90.

[4] He, X., Zhao, K., & Chu, X. (2021). AutoML: A Survey of the State-of-the-Art. Knowledge-Based Systems, 212, 106622.

[5] Truong, A., Walters, A., Goodsitt, J., Hines, K., Bruss, C. B., &Farivar, R. (2019, November). Towards automated machine learning: Evaluation and comparison of AutoML approaches and tools. In 2019 IEEE 31st international conference on tools with artificial intelligence (ICTAI) (pp. 1471-1479). IEEE.

[6] Nie, G., Rowe, W., Zhang, L., Tian, Y., & Shi, Y. (2011). Credit card churn forecasting by logistic regression and decision tree. Expert Systems with Applications, 38(12), 15273-15285.

[7] https://www.mdpi.com/2076-3417/10/1/90

[8] https://dl.acm.org/doi/abs/10.1145/3373509.3373578

[9] https://ieeexplore.ieee.org/abstract/document/8882834

[10] https://www.sciencedirect.com/science/article/abs/pii/S0950705120307516

[11] https://www.ml4aad.org/automl/

[12] https://www.alibabacloud.com/blog/6-top-automl-frameworks-for-machine-learning-

[13] applications-may- 2019_595317

[14] https://www.altexsoft.com/blog/datascience/machine-learning-project-structure-stages-roles-and-tools/

[15] https://isg.beel.org/blog/2020/04/09/list-of-automl-tools-and-software-libraries/

[16] https://analyticsindiamag.com/hands-on-tutorial-on-automatic-machine-learning-with-h2o-ai-and-automl/

[17] https://towardsdatascience.com/artificial-intelligence-made-easy-187ecb90c299

[18] https://www.kdnuggets.com/2020/01/h2o-framework-machine-learning.html

[19] https://machinelearningmastery.com/auto-sklearn-for-automated-machine-learning-in-python/

[20] https://autokeras.com/tutorial/structured_data_regression/

[21] https://machinelearningmastery.com/autokeras-for-classification-and-regression/