PREDICTION OF EMPLOYEE ACCESS CONTROL PERMISSION USING CATBOOST ALGORITHM

[1]Mrs.Pachala.Sunitha, [2]Sindhura Jannu, [2]Karri Jayanth Sai,
[2]Pinnamaneni Nithin, [2]Mohammad Fazallulah

[1]Assistant Professor (Ph.D), [2]Bachelor of Technology
[1]Computer Science & Engineering, [2]Computer Science & Engineering,
[1]Dhanekula Institute of Engineering and Technology, Ganguru, India

Abstract:In the corporate society, although timely information transmission is important for businesses, it is sometimes overlooked. We propose a conceptual framework for understanding delays in information transmission have a detrimental influence on employee outcomes in this project. Delays are viewed as workplace events using affective events theory. In contrast, to delay research, we investigate that workers experiences during a delay from temporal viewpoint damages, and these experiences impact interpersonal behavior too. We propose that employees' perception and reaction to delays is heavily influenced by their co-worker's conduct during the wait. We uncover a collection of situational and dispositional variables that are essential for forecasting delays in information transmission sabotage colleague relationships. Hence, employee access becomes a challenge across the corporate world and it is a chain reaction in which the whole task which is a subtask of a huge task gets delayed. Here in this project, we aim to predict if access is required for a particular resource request based on an indirectly interdependent parameter in an organization using Cat boost Classifier Neural Networks and compare that with XGB, Random Forest, Logistic Regression.

Keywords-Unbalanced Dataset, Categorical Feature, Logistic Regression, Random Forest, XGB, Cat-boost algorithm.

1.Introduction

There is a huge amounts of data and information with the role of workers within the company and its resources through which employees have core access. This information is useful for automating the core access permissions of employees within an enterprise.The core idea is generally to replace the process that is manual through utilizing the model of machine learning that is well trained utilizing the existing data which generally contain in-detail theory regarding the different attributes of the employees. Here these challenges can be mitigated/reduced by the cat-boost model, considering ML models appropriately. This particular model will help in order to automatically revoke or grant the core access /reduction of human involvement in the process. Feature engineering has also been considered as the core process in order to extract the features from raw data which are used to develop the core performance of algorithms. In this process we choose cat boost algorithm as the only solution, based on gradient boosting upon the decision trees. It is successor of core algorithm matrix net which is widely utilized within the enterprise for the purposes such as forecasting, making recommendations and ranking tasks. This is quite universal as well as will be easily applied across the broad range of the research areas to the variety of challenges.

2. Data Description

The data is taken from Amazon employee access with training set and testing set. We can observe from the table that each of the data samples has one label attribute called 'ACTION', where value '1' indicates this request is approved and "0" indicates rejection. In table, each samples has 8 features, which usually defines different role or group of one Amazon employee.

Feature Name	Feature Meaning
ACTION	"1": allow; "0": deny
RESOURCE	Resource Id
MGR Id	Employee manager Id
ROLE ROLLUP1	Organization role category Id1
ROLEROLLUP2	Organization role category Id2
ROLEDEPTNAME	Representation of Department
ROLETITLE	Business title illustration
ROLEFAMILYDESC	Role family put on illustration
ROLEFAMILY	Role family representation
ROLECODE	Unique Id for each Organization role

For the analysis of model, Receiver Operating Characteristic (ROC) curve is used to summarize classifier performances above trade offs between true positive and false positive error rates. We choose Area Under the ROC Curve (AUC) as a performance metric for imbalance classification problems. Since we have imbalance dataset, we use Matthews correlation coefficient (MCC) as another evaluation metric.

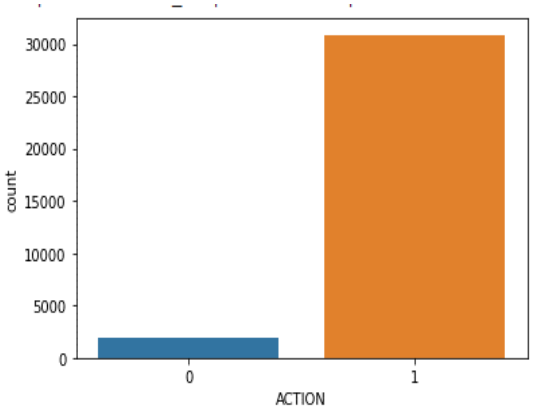
3. Methodology

3.1 **Data Preprocessing**

In this section,we divided this step into three parts: exploration of data, balancing the dataset, feature selection.

1.Data Exploration and Analysis:
Developing a deep understanding of various types of data is a significant requirement for the conduction of exploratory data analysis and feature engineering for machine learning. Most of the data can thus be grouped into four different types of machine learning: categorical data, text, numerical data, and time-series data.Numerical data- it refers to the data wherethe data points are exact numbers. This data has meaning as the data for measurement like house prices or counts. It can also be categorized by continuous or discrete data. Continuous data tends to assume any value within a range where different data have different values.Categorical data- categorical data tends to represent the features and can take numerical values. Categorical data is thus considered as the class label. It will be something as a person, or the property is either residentialor commercial (Tschang and Mezquita, 2020).Time series data- a series of numbers organized at regular intervals over a certain period of time. It is crucial particularly in fields or disciplines like finance.
In order to generate an effective result from the dataset, we use python programming language and import multiple libraries such as seaborn, NumPy, pandas and warnings. The copy() function plays a vital role in copying the dataset, and an info() function helps get the information regarding the data. It identifies the data type of each column; for example, the datatype of mrg_id is int64.nunique() is one of the most important functions of python that returns unique elements.
2. Imbalanced data sets:

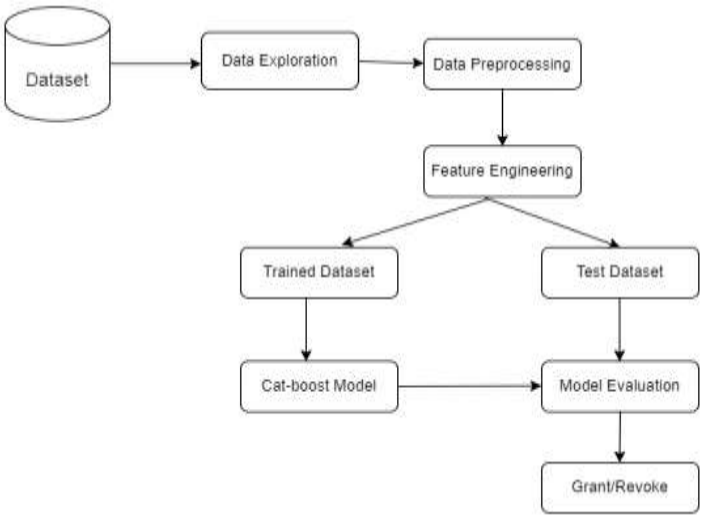
It is important to incorporate the countplot() function into the dataset to create a graph of the resources. Here the function contains two arguments, such as x and data. X indicates 'ACTION,' and data considers data_explore. After analyzing the source code and result section, it can be said that the data_explore array incorporates the group by() function and count() function. The sort_values(), head(), and transpose() functions are included in the coding section in order to get the number of actions of each resource. The output section indicates that resource 4675 contains 836 actions. It means resource 4675 has been accessed 836 times.



3.Feature Selection:In this dissertation we consider an imbalanced dataset that incorporates information regarding multiple records. The heat map or correlation matrix play's a crucial role in comparing several features of dataset. This matrix includes action, resource, Mrgid, Rolerollup, Roledeptname, Roletitle, Rolefamilydesc, Rolefamily and Rolecode. This matrix identifies the correlation between two variables; for example, the values of the diagonal cells are 1.Some variables are positively correlated with other variables, and other variables are negatively correlated. Action is one of the most important variables of the correlation matrix, which is positively correlated with the resources, Rolerollup, Rolefamilydesc, Rolefamily, and Rolecode.

The resource is another variable of the correlation matrix that is negatively correlated with rolerollup.

3.2 System Architecture:



4. FORECASTING MODELS:

In this module we briefly discuss about few important algorithms used in the process of modelling.

1.Logistic regression- used for estimating discrete values, and is dependent on independent variables. In general, probability of occurrence can be forecasted by fitting data into a logit function. It has also been named logit regression. Other than that, it also forecasts the probability, and the values of output lie between 0 and 1. There are also various steps for developing the model like regularization techniques, interaction terms, non-linear model, etc.

2.Decision tree- is one of supervised learning algorithm which is used mostly for classifying various kinds of problems. It functions for both categorical and continuous dependent variables. In this algorithm, the population is split into two or more similar set (ZuiderveenBorgesius, 2020). It is done depending on most of the attributes that are significant or independent to create as distinct groups as possible.

3. Random forest- It is the trademarked term for resembling a tree structure. In random forests, the decision trees are collected and are referred to as 'forests'. For the classification of a new object that is dependent on attributes, each tree classifies a new objects that is based on the attributes. The forest then selects the classification having the maximum number of votes.

4. GBM- It is a boosting algorithm that is used to deal with lot of data for making a prediction with high power of forecasting. It is an ensemble of learning algorithms that merges the prediction level of several best investors for developing robustness over a single estimator. It also tends to combine weak and average predictors for setting up strong predictors. It always competes with AV Hackathon, CrowdAnalytix, and Kaggle.

5.XG Boost is a classic gradient boosting algorithm that is famous for winning and losing some Kaggle competitions is XG Boost. It has extremely high predictive power that tends to make the best choice for accuracy and possesses both tree learning algorithm and linear model, and the algorithm becomes faster in this case with the existing gradient boosting techniques. It tends to support various functions of the objectives like classification, ranking, and regression. One Of the biggest advantages of XGBoost is that it is a regularized boosting technique, and supports various languages like Java, C++, Python, and so on.

6. Catboost- is a recently open-source machine learning algorithm from “Yandex”. It can integrate easily with deep learning frameworks like Apple’s Core ML, Google’s TensorFlow, and many more. The finest part about this algorithm is that it does not need extensive training of data like other ML models, and is not functioning on a variety of formats, not undermining how robust it is.The above-mentioned algorithms are the most common algorithms that are used. In this stage, the process of setting up the algorithms will be discussed.The aspects which are required to setting up the suitable algorithm which will be best for the situation are known by following the below mentioned steps:

The first step is to categorize the problem, and it is to be categorized both by input and output. Then, understanding the data is important, the process also involves analyzing, processing, and transforming data. The third step is to search for the available algorithms, then, the machine learning algorithms are to be implemented. The fifth and last step is to optimize the hyperparameters.

4.1 Employee Access Data Prediction Using CAT-BOOST Algorithm

The Cat-Boost algorithm can be applied to employee data to analyze the data. The analysissection clearly shows that the developers need to import multiple packages such as pandas,seaborn, cat-boostclassifier, etc. After importing all necessary packages, uploading the employeedata using the read_csv() function is adequate. In order to encode the categorical variables, it is adequate toincorporate feature engineering.This machine-learning algorithm also divides the datasetsinto two parts, such as test data and train

data. After this step, the algorithm plays a crucial rolein creating the models, which helps to forecast the employee data. It is essential to understandthe model's performance by considering precision, recall, and accuracy score. After analysing multiple research papers on the Cat-Boost model, it can be said that the accuracy level of thismodel is 98.55%. Thisalgorithm supports the operations or the activities which are involved with the process ofrunning out of the box along with proper classification as well as accurate regression.

Importance of CatBooset for collecting the employee data:

The CatBoost algorithm is beneficial for the particular scenario because it is highly accurate for the model building along with a great graphic processing Unit.

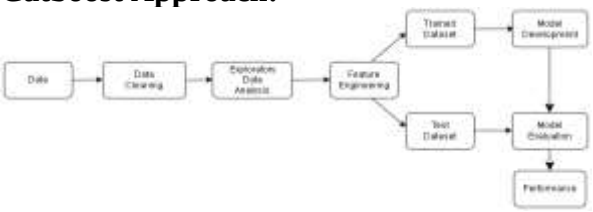
The Catboost model gives the researcher remarkable results along with the default parameter.In this dissertation, the author uses the categorical variable, which is effective for the CatBoost algorithm; there is no need to preprocess the variable through this procedure (Hamilton and Sodeman, 2020). The CatBoost algorithm provides excellent visualization such as training process and feature importance. For the python package, the CatBoost algorithm is very effective.

Advantages of CatBoost to grab the employee data

The first advantage of the Catboost is Robust; with the help of the CatBoost model, it is easy to improve the performance of the other model while decreasing the overfitting as well as time. Catboost has different parameters for numerical data, but it reduces the hyper-parameter number because the default parameter develops a great result (Song and Baicker, 2019). When accessing the employee data, it is understood that the CatBoost algorithm is high performance and a greedy novel that is effective in implementation. So that in the competitive market, the CatBoostalgorithm is very effective.

From the analysis of different research papers, it is understood that the CatBoost algorithm is one of the better algorithms than the XGBoost algorithm. When cleaning the data, the data is automatically converted into numerical data so that the machine can easily understand and predict the upcoming data. When any text-related model converts into numerical data, the CatBoost algorithm is effective in this situation. This scenario uses numerical data to grab the employee data. For this reason, the author uses the CatBoost model to extract the data.The main reason to choose the CatBoost technology is that it is very efficient and efficiently works with non-numeric data.

Catboost Approach:



Citation: Hamilton and Sodeman, 2020

From the above image, it is identified that the procedure of the CatBoost algorithm is helpful to solve the regression problem of the employee data. The first process that is understood from the image is collecting the data, then cleaning the data, data analysis, and using feature engineering. Feature engineering has two different types of features: Train Dataset and Test Dataset. These two features play an essential role in Model development as well as Model Evaluation.

4.2 IMPLEMENTATION AND RESULTS:

Logistic regression to analyze employee data:

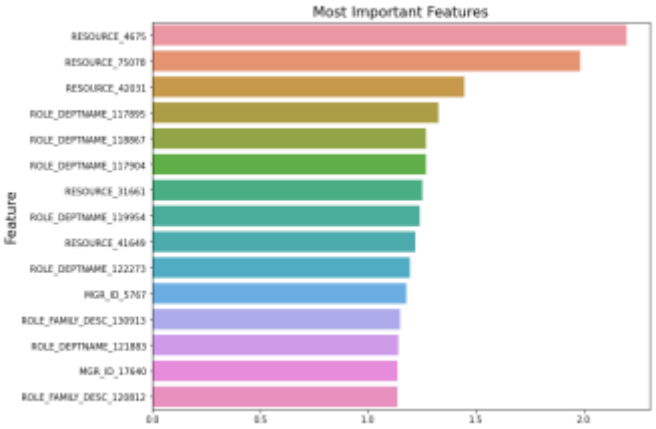
From the result and analysis section of this dissertation, it is clear that the logistic regression algorithm plays a vital role in analyzing historical data of employees. Logistic regression is a classification algorithm that analyses employee historical data to predict employee resource access.

The logistic regression model contains four essential steps: data processing, feature engineering, model building, and model evaluation. In the data processing step, the python developers need to import three libraries such as pandas, seaborn, matplotlib.pyplot. In order to get an overview of the data, it is essential to use the describe() function. The data processing step of the logistic regression model plays a vital role in handling missing values which enhances the usability of the dataset. The authors of this dissertation use the IsNull() function to check if the value is null or not. It is also essential to calculate the percentage of null values in a column for further processing. It is practical to apply both IsNull() and count() functions to calculate the number missing value. In order to create a data frame, developers use the dataframe() function . If a column contains many missing values, it is better to drop the column to conduct an effective prediction. Feature engineering is another crucial step of the linear regression model, which plays a crucial role in analyzing the historical data of employees. There are four steps present in address outliers, feature transformation, categorical feature encoding, and

feature selection. After analyzing a research paper, it can be said that logistic regression accepts a numeric value. The categorical data need to be converted from categorical data to numbers. There are two types of techniques present for data encodings such as label encoding and one-hot encoding. The label encoding is applicable for the dataset where data contains high cardinality, and one-hot encoding is appropriate for low cardinality data. It is practical to select label encoding for analyzing the historical data of employees. After conducting the label encoding, it is essential to transform the data into integer and float—the authors of this dissertation use this encoding process for generating int64 and float64 values.

Feature selection incorporates correlation analysis in order to identify highly correlated variables. It helps to develop the correlation matrix by understanding the relationship between each variable. Model building is the next step of the logistic regression algorithm, which plays a vital role in prediction. In this step, it is essential to divide the variables into two parts, one is independent variables, and another one is dependent variables (Helbich*et al.*, 2020). The independent variables indicate input features, whereas the dependent variables consider labels. After analyzing the result section and some research papers, it can be said that features and labels are divided into two subparts. Here one subpart is for testing, and other subparts help with training. The train_test_split() function is one of the most critical functions in this step which needs to incorporate the text size parameter. Model evaluation is the next part of the linear regression model that incorporates four elements evaluating the employee dataset: accuracy, confusion matrix, AUC, and ROC. The evaluation process generates matrices that calculate the difference between test values and predicted values. It can generate four types of results: True Positive, True Negative, False Positive, and False Negative. In order to generate the confusion matrix, it is practical to use plot_confusion_matrix(), which provides an appropriate visual representation. It is essential to get a high accuracy value for employee data to enhance the performance of the employee resource access system (Helbich*et al.*, 2020).

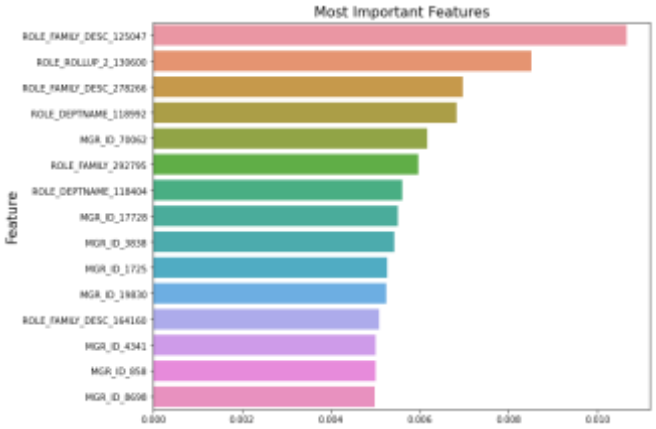
Features obtained by Logistic regression



XGBoost algorithm to analyze employee data:

XGBoost algorithm is a very effective machine learning algorithm that can be applied to historical data of employees to predict employee resource access. It is a decision Tree-Based machine learning algorithm that incorporates a gradient boosting framework. The prediction problem of the employee dataset contains unstructured data such as text. Apart from unstructured data, the employee dataset also incorporates structured data. After analyzing a research paper, it can be said that the XGBoost algorithm can be applied in different processes in the dataset. It helps to solve classification, regression, user-defined prediction, and ranking problems. The employee resource access problem comes under the classification problem, so using the XGBoost algorithm for data analysis is adequate. This algorithm can be used smoothly on multiple operating systems such as Linux and windows. Developers can use multiple programming languages such Julia, Scala, Java, C++, Python, R, etc. In this dissertation, the authors select the Python programming language for implementing the XGBoost algorithm (Nocker and Sena, 2019). This machine-learning algorithm conducts algorithmic enhancement and optimization. Developers use this algorithm for analyzing employee data because of six reasons. The XGBoost algorithm avoids over-fitting and efficiently handles missing data. Apart from these, it also has cross-validation capability and cache awareness.

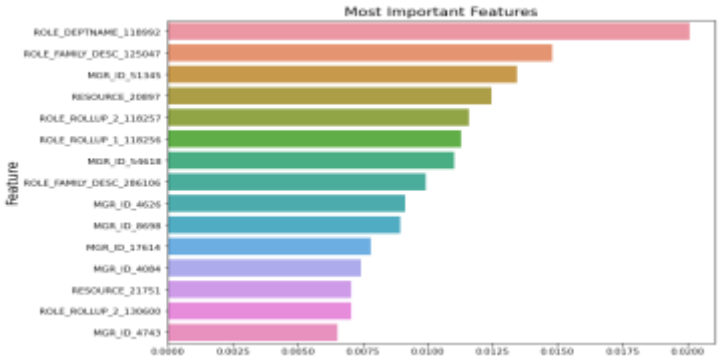
Features obtained by XG Boosting



Random Forest to analyze employee data:

Random forest is the trademarked term for resembling a decision tree. In random forests, the decision trees are collected and are referred to as 'forests'. Forthe classification of a new object that is dependent on attributes, each tree classifies anew object that is based on the attributes. The forest then selects the classification having the maximum number of rating/votes. Random forest algorithm is one of the most effective machine learning algorithms. It is effective to collect historical employee data and conduct classification through random forest algorithms. The random forest classifier considers discrete label data, and it plays an important role in improving the accuracy level of the prediction. The random forest algorithm decreases over fitting and plays aimportant role in improving accuracy. This machine-learning algorithm is flexible for both Regression and Classification problems. In the employee data analysis process, the random forest algorithm helps to automate missing values. It is not important to normalize data because it follows a rule-based approach. It is effective to import the NumPy library as well as the pandas package. The NumPy library works with arrays and plays a vital role in managing matrices, Fourier transforms, and linear algebra. The pandas package incorporates the python code to design expressive, flexible, and fast data structures. In order to collect historical data of employees, it is important to import a. CSV file. Read_CSV() is an effective function of python which helps to study data from the CSV files. After uploading the dataset, it is important to split the data set into two parts: Training data and Testing the dataset. The training part of the dataset contains 75% data, whereas the testing part of the data includes 25% data. In order to standardize the data, it is effective to incorporate Standard Scaler. The fit_transform() function considers train data and transform() takes test data. After fitting and predicting the data, it is important to evaluate the confusion matrix. The confusion_matrx() function considers two arguments, one is a test, and another one is a prediction.Thefit_transform() function considers train data and transform() takes test data. After fitting and predicting the data, it is important to evaluate the confusion matrix. The confusion_matrx() function considers two arguments, one is a test, and another one is a prediction. The visualization of training data and testing data sets help to analyze the historical data of employees.

Features obtained by Random Forest:



Cat Boost to analyze employee data:

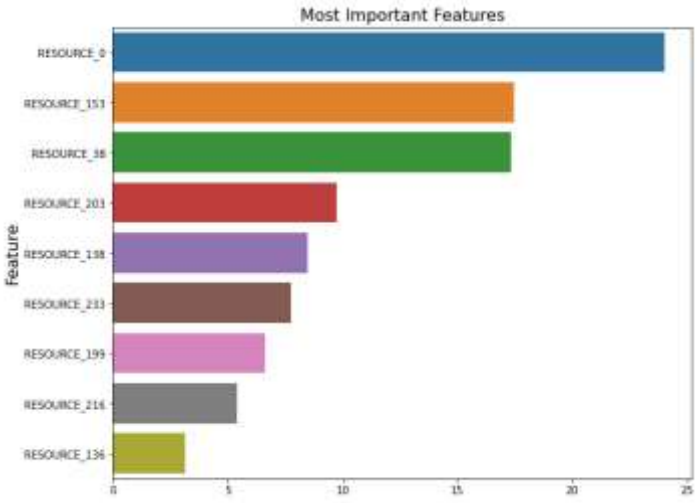
The CatBoost algorithm can be applied to employee data to analyze the data. The analysis section clearly shows that the developers need to import multiple packages such as pandas, seaborn, catboostclsifer, etc. After importing all necessary packages, uploading the employee data using the read_csv() function is adequate. The head() function plays a vital role as a header to all columns. In order to encode the categorical variables, it is adequate to incorporate feature engineering. This machine-learning algorithm also divides the datasets into two parts, such as test data and train data. After this step, the algorithm play's animportant role in creating the model, which helps to forecast the employee data. It is essential to understand the model's performance by considering precision, recall, and accuracy score. After analyzing multiple research papers on the CatBoost model, it can be said that the accuracy level of this model is 98.55% (Pulicherla*et al.*, 2019). In this case, the Catboostalgorithmcan be considered to be very useful as it provides the necessary assistance in terms of achieving best results on the benchmark. Within the employee data, it is observed that the categorical features play a crucial role and the Catboost algorithm can significantly improve those features along with making them very significant as well as undeniable. While working with the employee data, this particular algorithm can successfully enable faster predictions. The default parameters of this algorithm are considered to offer a better starting if compared to the

GBDT algorithms. Some of the most effective and major advancements of the Catboost algorithm are object importance, features interactions as well as the snapshot support. This algorithm supports the operations or the activities which are involved with the process of running out of the box along with proper classification as well as accurate regression. This algorithm introduces two major algorithmic advances. The first one is regarding the process of successful implementation of the ordered boosting.

Features obtained by Cat Boost:

On the other hand, the second algorithmic advance is the permutation-driven alternative to the innovative algorithm along with the classic diagram in order to successfully execute the process of processing the categorical features.

Both of those effective techniques which are associated with the Catboost algorithm, utilize random permutations of the training such as fighting the prediction shift which is caused by the special kind of target leakage. So, it can be successfully ensured that the Catboost algorithm provides the necessary assistance to improve the process of gathering as well as accessing employee data.



Voting Classifier Model:

In order to train the dataset, it is important to import the Voting Classifier model. It helps to train employee resource datasets and plays a vital role in prediction. It supports two types of voting systems, one is soft voting and another one is hard voting. In this scenario, the authors use soft voting to get the highest probability result.

5.MODEL EVALUATION:

In this section, we identify the accuracy level of different machine learning algorithms. The coding analysis identified that the logistic regression accuracy level is 85.5, the Random forest accuracy level is 80.5, the XGBoost accuracy level is 84.3, and the last is CatBoost which accuracy level is 89.2 So, it is previously proof that a high accuracy level algorithm is effective for employee data analysis.

Performance Metrics

In these paper we have two performance metrics ROC(Receiver Operation Curve) ,AUC(Area Under Curve) and also we are using MCC (Matthews correlation coefficient). Value of MCC is lies between -1 to +1. The coefficient value of +1 represents a perfect prediction, 0 represents average random prediction and -1 an inverse prediction.MCC value will be high only if model has high accuracy on predictions of negative data instances as well as of positive data instances.

${MCC} = {(TP + TN) * (FP + FN)}/{\sqrt{((TP +FP) * (TP + FN) * (TN + FP) * (TN + FN))}}$

	Model	CV Train AUC Score	CV Test AUC Score	CV Train MCC	CV Test MCC
1	Logistic Regression	0.847	0.752	0.293	0.184
2	Random Forest Classifier	0.817	0.744	0.000	0.000
3	XGB Classifier	0.836	0.725	0.391	0.227
4	Cat-Boost Classifier	0.882	0.778	0.490	0.258
5	Voting Classifier	0.869	0.765	0.300	0.173

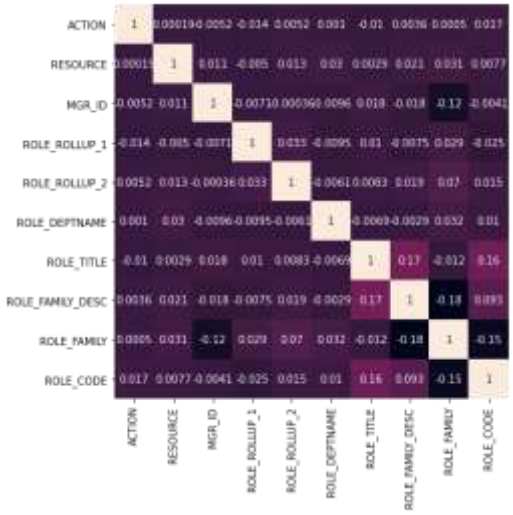
Accuracy = (TP+TN)/(TP+FN+FP+TN)

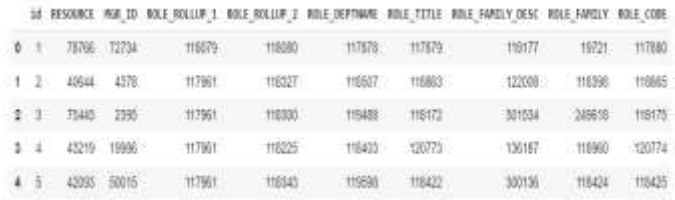
Precision (PR) = TP/(TP+FP)

Recall (DR) = TP/(TP+FN)

F1-score = 2*{(PR*DR)/(PR+DR)}

MCC -Correlation matrix/heat map of the features compared to each other:

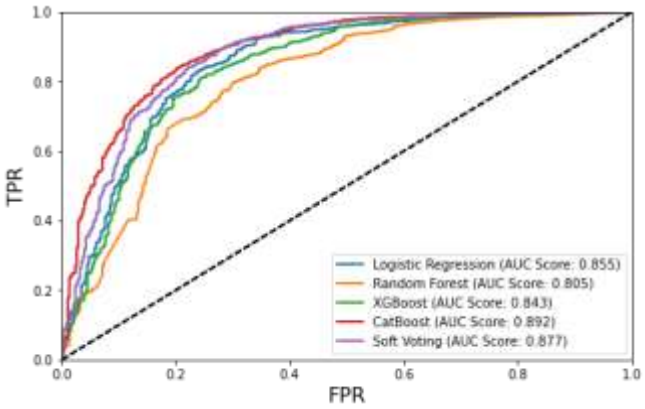




	id	RESOURCE	MGR_ID	ROLE_ROLLUP_1	ROLE_ROLLUP_2	ROLE_DEPTNAME	ROLE_TITLE	ROLE_FAMILY_DESC	ROLE_FAMILY	ROLE_CODE
0	1	78766	72734	118079	118080	117878	117879	118177	19721	117880
1	2	40644	4378	117961	118327	118507	118863	122008	118398	118865
2	3	78443	2395	117961	118300	118488	118172	301534	248618	118179
3	4	43219	19986	117961	118225	118403	120773	136187	118960	120774
4	5	42093	50015	117961	118343	119598	118422	300136	118424	118425

	id	ACTION
0	1	1
1	2	1
2	3	1
3	4	1
4	5	1

ACCURACY SCORE GRAPH OF VARIOUS ALGORITHMS:



False Positive rate-FPR = FP/(FP + TN)

True Positive rate-TPR = TP/(TP + FN)

From the analysis of the accuracy level and the ROC level, it is understood that Catboost is a more effective algorithm than the other algorithm. It is also good to find out the MCC value. So the author takes the final decision to choose the CatBoost algorithm for the employee data analysis. The test_data_head() function makes it easy to predict the value from the employee data set. The output of the screenshot shows ID, resource, MGRID, ROLEROLLUP1, ROLEROLLUP2, RoleDeptName, ROLETITLE, ROLEFAMILYDESC, ROLEFAMILY, ROLECODE. Through this procedure, it is easy to grab effective data from the output. Also, the output shows that the CatBoostalgorithm is effective for this particular situation.

6. Future Enhancement:

From this paper which is based on theamazon employeedata,we haveshown an automation of systems to accessdata using cat-boost algorithm for imbalanced dataset and predicted a best accuracy compared to other algorithms. In future we can use many other algorithms and also we can extend create web pages like login etc to this work.

7. CONCLUSION:

In these paper we had choosen an imbalanced dataset from the amazon employee access to provide an automatic access to the employee without human involvement using AI solution. Here in these paper we had used cat-boost algorithm for providing bestaccuracy compared to other algorithms using imbalanced dataset. We used four algorithms Logistic Regression, Random Forest, XG Boosting and Cat-boost algorithms to find the performance metrics of both ROC and AUC. We had also used MCC for evaluation. The cat-boost algorithm it provides 89% accuracy for imbalanced dataset compared to other algorithms.

REFERENCES:

[1] Aodi Liu, Xuehui Du, and Na Wang“Efficient Access Control Permission Decision Engine Based on Machine Learning”, Information Engineering University,Zhengzhou,450000,China,Volume2021,ArticleID3970485 https://doi.org/10.1155/2021/3970485

[2] Shijian Tang, Jiang Han and Yue Zhang “Amazon Employee Access Control System”, Department of Electrical Engineering, Stanford University,https://cs229.stanford.edu/

[3]FangrongZhou,Hao Pan, Zhenyu Gao, Xuyong Huang,[1] Guochao Qian,[1] Yu Zhu,[2] and Feng Xiao[3]"Fire Prediction Based on CatBoost Algorithm", Volume 2021 ArticleID 1929137 https://doi.org/10.1155/2021/1929137

[4]Saddam Hussain ,Mohd. Wazir Mustafa ,Touqeer A. Jumani b, Shadi Khan Baloch , Hammad Alotaibi , Ilyas Khan ∗ , Afrasyab Khan f,"A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theftdetection",2021 Published by Elsevier Ltd.

[5]Badirli, S., Liu, X., Xing, Z., Bhowmik, A., Doan, K. and Keerthi, S.S., 2020. Gradient boosting neural networks: Grownet. arXiv preprint arXiv:2002.07971.

[6]Cheng, L. and Wang, H., 2021. CatBoost model with synthetic features in application to loan risk assessment of small businesses. arXiv preprint arXiv:2106.07954.

[7]Zamri, N.E., Mansor, M., MohdKasihmuddin, M.S., Alway, A., MohdJamaludin, S.Z. andAlzaeemi, S.A., 2020. Amazon employees resources access data extraction via clonal selection algorithm and logic mining approach. Entropy, 22(6), p.596.https://www.mdpi.com/1099-4300/22/6/596/pdf

[8]Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N. and Asadpour, M., 2020. Boosting methods formulti-class imbalanced data classification: an experimental review. Journal of Big Data, 7(1),pp.1-47.

[9]Pulicherla, P., Kumar, T., Abbaraju, N. and Khatri, H., 2019, May. Job shifting predictionand analysis using machine learning. In Journal of Physics: Conference Series (Vol. 1228,No. 1, p. 012056). IOP Publishing. https://iopscience.iop.org/article/10.1088/1742-6596/1228/1/012056/pdf

[10]Gao, X., Wen, J. and Zhang, C., 2019. An improved random forest algorithm for predictingemployee turnover. Mathematical Problems in Engineering, 2019.https://www.hindawi.com/journals/mpe/2019/4140707/

[11]Beam, A.L., Manrai, A.K. and Ghassemi, M., 2020. Challenges to the reproducibility ofmachine learning models in health care. Jama, 323(4), pp.305-306.https://www.ncbi.nlm.nih.gov/pmc/articles/pmc7335677/

[12]Li, K., Mai, F., Shen, R. and Yan, X., 2021. Measuring corporate culture using machinelearning. The Review of FinancialStudies,34(7),pp.3265-3315.https://www.researchgate.net/profile/Feng_Mai2/publication/328471762_Measuring_Corporate_Culture_Using_Machine_Learning/links/6035bcb4299bf1cc26e7e95c/Measuring-Corporate-Culture-Using-Machine-Learning.pdf

[13]Peñalvo, F.J.G., Benito, J.C., González, M.M., Ingelmo, A.V., Prieto, J.C.S. and Sánchez,R.T., 2018. Proposing a machine learning approach to analyze and predict employment andits factors. IJIMAI, 5(2), pp.39-45. https://documat.unirioja.es/descarga/articulo/6907747.pdf

[14]Ibrahim, A.A., Ridwan, R.L., Muhammed, M.M., Abdulaziz, R.O. and Saheed, G.A., 2020.Comparison of the CatBoost Classifier with other Machine Learning Methods. InternationalJournal of Advanced Computer Science and Applications, 11(11).

[15] Zeadally, S., Adi, E., Baig, Z. and Khan, I.A., 2020. Harnessing artificial intelligence capabilities to improve cybersecurity. *Ieee Access*, *8*, pp.23817-23837.Available at:https://ieeexplore.ieee.org/iel7/6287639/8948470/08963730.pdf

[16] Song, Z. and Baicker, K., 2019. Effect of a workplace wellness program on employee health and economic outcomes: a randomized clinical trial. Jama, 321(15), pp.1491-1501. https://jamanetwork.com/journals/jama/articlepdf/2730614/jama_song_2019_oi_190030.pdf

[17] Futia, G. and Vetrò, A., 2020. On the integration of knowledge graphs into deep learning models for a more comprehensible AI—Three Challenges for future research. *Information*, *11*(2),p.122.availableat:https://www.mdpi.com/2078-2489/11/2/122/pdf

[18] Jagtiani, J. and Lemieux, C., 2019. The roles of alternative data and machine learning in fintech lending: evidence from the LendingClub consumer platform. *Financial Management*, *48*(4), pp.1009-1029.https://leedsfaculty.colorado.edu/bhagat/FintechLending.pdf