Original Article

Prognosticative Approach For Intensifying E-Commerce And Pharmaceutical Industry With Artificial Intelligence In Cybernetics

Prof (Dr). K.P.Yadav[1*], Dr. Sandeep Kulkarni[2]

[1*]Presently Vice Chancellor MATS University, Raipur. BTech, Mtech, PGDBM, PhD, Post Doc Fellow, D.sc (Doctor of Science), Email- drkpyadav732@gmail.com
[2]Presently working as software professional, Bangalore. M.sc, PhD, Email- sandeeppostdoc@gmail.com

***Corresponding Author:-** Prof (Dr). K.P.Yadav

*Presently Vice Chancellor MATS University, Raipur. BTech, Mtech, PGDBM, PhD, Post Doc Fellow, D.sc (Doctor of Science), Email- drkpyadav732@gmail.com
DOI:10.47750/pnr.2022.13.S08.507

Abstract

Every monthly event takes place in super market attached with pharmaceutical store with the single database attached in Bangalore city. The mall which we are working as the case study people around India, as Bangalore is the IT capital. As people stay far away from family, usually it is necessary to take all the things for the daily needs including grocery, pharmacy, ration and other commodities. As we know that India is a diverse country. People here in India has different ethnicity which is indirectly related to different language speaking population. Here we are making use of data to increase the sale of the store.

Keywords: Machine learning, Logistic regression, KNN, MLP, XGBoost, LSTM, Accuracy, Confusion matrix

INTRODUCTION

To increase the sales of the store, we are taking all the people's data who are working with the store, who has the data of the customers, frequently visits the store and purchase the items from the store. We collected the data of last 3 months data of the customers which includes where the customer belongs originally or native place the customer belong. We took the data from the staff after that we preprocessed the data, applied machine learning algorithms and took accuracy of the data with other confusion matrix for the classification task. So here we are classifying the different language speaking persons. So that once we segregate this thing it becomes easy for the store manager to sell the specific and recommend the products for the customers.

AIM OF THE STUDY

Motive of the research paper to get the profit in ecommerce business with is attached to pharmacy store using common database to evaluate the quality of business features such as customer gender and other attributes of the customer.

MATERIALS AND METHODS

Study Settings: The study was conducted in store at Malleshwaram, Bangalore.

Sample Size: 2000 customers are around Bangalore.

FEATURE VARIABLES

Customer details that is address, age and other details staying in the Bangalore.

Administration and Ethics

The Study was approved by MATS University in the department of Computer Science.

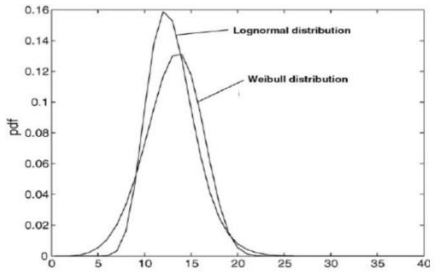
Ratio Study

This was conducted on 20% of customers who visits the shopping mall on monthly basis.

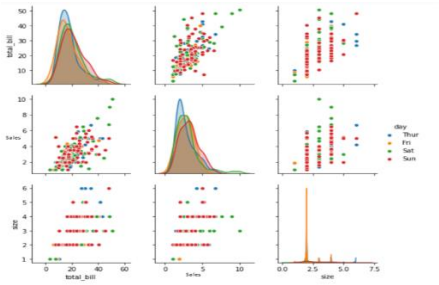
Procedure

Before starting the process we started to do the preprocessing steps as we took all the billing information which is present such as pricing, products. We took the all the information basically it was in PDF format. We used the tessaract library to get the information from the PDF. We got all the information from it specifically information like gender, name, language, address details from the billing information. We took all the billing information that are needed for building the

classification model. We considered the details as the features of the data. Imported all the tessarat to comma separated files. After importing the files then we did data cleaning , there was some cases that customer's name or address or gender or language was missing , so we used we try to fill the missing values using imputation technique. So making sure that it does effect the model while implementing in machine learning technique. So finding the outlier if there any outlier we are ignoring the value by squashing mechanism. Lower casing helped in the making sure that all the char are in same case. Removed stopped words, so that unnecessary words are making noise in the models. After that we tokenized the data by using NLTK. By using the probability distribution function we wanted to find the language distribution, how many people are there in this particular region or area which the store is there, now we came to know that data spread across is normal distribution for approximately 3000 customers. As we normalized the data value fits in 0 to 1. As most of the data are lognormal distribution, we find the average age for kids section to be 15 years old and price they purchase be around same price.



Probability distribution function for Sales
Age Verses products points



STUDIES AND FINDINGS

As we are working on text data, it is not possible to use the words, we need to have some form so that the machine learning model understands so we used TFIDF vectorizier most of its value about 0.049 for each words. So that we come to know that frequency of customer information which is obtained. We analyzed the gender and language, we thought it was important because ladies buy different products as compared men buying different products in both super market and pharmacy arena, how they have been correlated using plot. As there are many fields in the billing receipt if we consider as the features, need to have the principle component analysis to reduce the dimensionality. Working on convert to vector we used word to vector format using bag of words as machine learning algorithms cannot understand raw English words so it was necessary to convert it to vector form with semantic similarity. As the case progressed it created confusion so using classification method to get the context meaning related to name, language. For instance if we are looking for the Kannada language customers, Tamil speaking customers have most similar taste as they are culture similarity so there might be chances that customers like same taste in buying the products as used K nearest neighbors algorithms after vectorization.

The angle between vectors ***u*** and ***v*** can be defined by

$$\cos\theta = \frac{u \bullet v}{\|u\|\|v\|}$$

The vectors are parallel if ***u*** and ***v*** are scalar multiples.
The vectors are perpendicular if $u \bullet v = 0$

$$x_{\text{stand}} = \frac{x - \text{mean}(x)}{\text{standard deviation}(x)}$$

TFIDF

For a term i in document j:

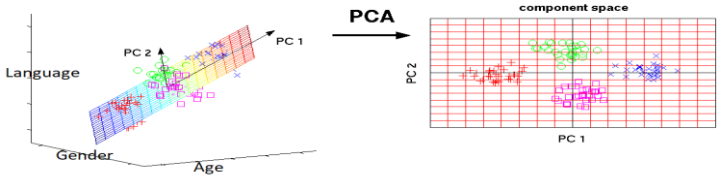
$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

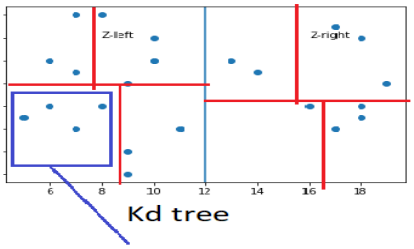
Where u and v are considered two points of the customer feature, finding the distance using the cosine theta, we considered the clusters which are near by the points with the help of cosine similarity and local sensitive hashing that are grouped together used for optimization process, made sure that there was no over or under fitting. As most of the columns are not in proper values we did column standardization. We used classifier known as logistic regression with the sigmoid function for the squashing so finding the right hyper parameter finding we took route of grid search method.

$$\text{Var}[a^T X] = \frac{1}{n}\sum_{i=1}^{n}\left\{a^T\left(X_i - \frac{1}{n}\sum_{j=1}^{n} X_j\right)\right\}^2$$
$$= a^T V_{XX} a.$$

Principle component analysis



So now used theanother support vector classifier so that the margin distance gets maximized for the points separation using the kernel trick and polynomial kernel. There were some border points and noise points removed that points. Used min points and epsilon points as hyper parameter. We reduced the feature using principle component analysis. We experimented with many k values for getting the mean error as minimal as possible.

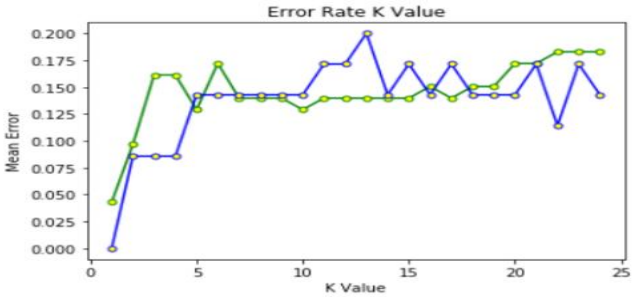


Kd tree

Bounding Box

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

KNN



Placed the equivalent group of clusters of points where the customers belong to that segments of points using kd tree based on that we figured out the KNN nearest neighbor points and putting that into that particular group with x as input with the right k of value 12 across the mean error with the values of x and y for n points. As this store has online and offline store as , even we have the online store , after the presence of machine learning algorithm we put it in to database.

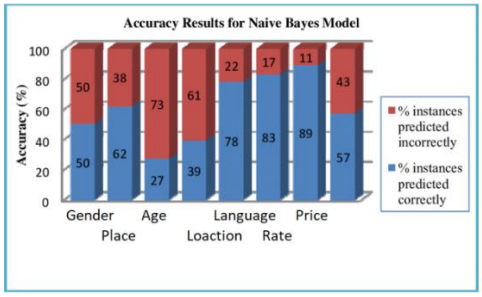
Machine and Deep learning algorithms experiments

Getting the values into excel value from the sql database. As database gets stored everything from the starting of the store. We need to get the recommendations products for the customers. Once we got thewhich all the products the customers could be using then will be clustering mechanism k means clustering so that all the clusters come across one group each. Local reachability distance of points make for all the points so that minimum points for the distance of all points so that using this we can minimize the outlier points.

$$P(\text{User} \mid \text{Positive}) = \frac{P(\text{Positive} \mid \text{User})P(\text{User})}{P(\text{Positive})}$$

$$= \frac{P(\text{Positive} \mid \text{User})P(\text{User})}{P(\text{Positive} \mid \text{User})P(\text{User}) + P(\text{Positive} \mid \text{Non-user})P(\text{Non-user})}$$

Bayes Theorem



We used to another classifier known as naïve bayes so that to know the probabilities of the values during the classification. We took the bayes classifier to classify the points based on the customer/user who are coming to store classified based on probabilities of languages. If the probabilities is more of that language people are staying near by the store. In the naïve bayes algorithms we go the predicted probabilities for the feature values.

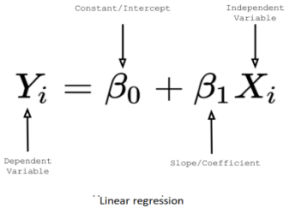
$$P = \frac{e^{a+bX}}{1 + e^{a+bX}}$$

Logistic regression

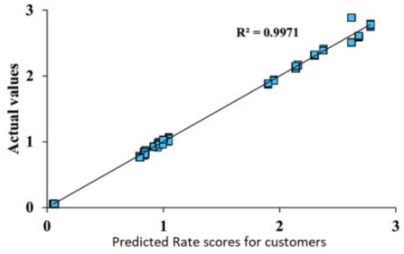
```
clf = LogisticRegression(penalty='l1', C=c, solver='liblinear')
clf.fit(X_train_mm, y_train)
```

```
Training accuracy: 0.9218200468933178
Test accuracy: 0.9249926750659244
```

For classification task based on the base location which people stay originally so called native place. We used another classifier logistic regression by this we classified different location with accuracy of 92%. In linear regression we took all the features of the dataset to predict the boundaries of the people based on the features. Accuracy of linear regression is 99%. By expansion of this ecommerce industry with the help of chat bot where people can use the chat bot as the tool , for instance if a customer is purchasing the product on online ,if he does not want to go to store, if he wants certain products from the store then the person is able to purchase the product based on the suggestion given by the AI robot, so that this might be the one of the reason that ecommerce business expansion plan can be executed with less number of man power involved in this business.



Linear regression



We used gradient decent for the variables x and y using linear regression gradient decent. While using support vector machine by lagrange's multiplier for optimization for classification of points x and y 's with the accuarcy of 99%. Now using the decision tree for classififying the people based on the ethnics we used the labels as language so that classification becomes easy with that motto with the higher accuracy . We got the information gain, gini impurity along with feature standardization. Whenever customer searches the product, it becomes very handy. Customer search results become very effective.

$$\sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (\mathbf{x}_i \cdot \mathbf{x}_j)$$

Dual form SVM

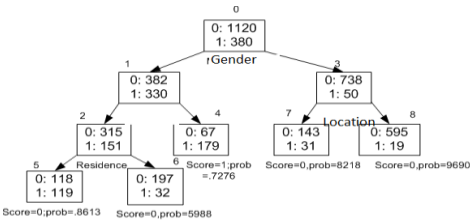
Accuracy Score: 0.8003240683036271

	precision	recall	f1-score	support
0	0.80	0.81	0.80	8033
1	0.80	0.80	0.80	8013

For classification task based on the language which people talk so called their native language. We used another classifier Support vector machine by this we classified different languages with accuracy of 80%. It internally ranks the product so that while searching the product based on the customer product requirement very fast. Suppose customer buys the product, similar products has been auto suggested for the customer. Smoothing the user experience had before AI powered system, this problem has been solved. Based on the customer capacity product has been suggested to customer ordered by pricing, so that every customer buys the based on the pocket size of the customer.

$$Gini = 1 - \sum_{i=1}^{C} (p_i)^2$$

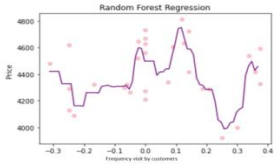
Decision tree



Random forest we knew the feature f importance value for each node n for to determine the label y. Along with it we did stacking of the model of Random forest , so that regression value so that more gradient boosting for cascading the classifiers which is scalable which is also good in parallel processing the data and it is more efficient with the accuracy of 96%. Based on the past behaviour of the customer like weather he or she buys what kind of products , algorithms bifericates and suggests and recommends the product which customer can buy it.

$$\frac{1}{N} \sum_{i=1}^{N} (fi - yi)^2$$

Random forest




```
#printing error data
print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, Y_pred1))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, Y_pred1))
print('Root Mean Squared Error:', np.sqrt(metrics.mean_squared_error(y_test, Y_pred1)))
```

```
Mean Absolute Error: 65.94614329034201
Mean Squared Error: 9219.704734824747
Root Mean Squared Error: 96.01929355512227
```

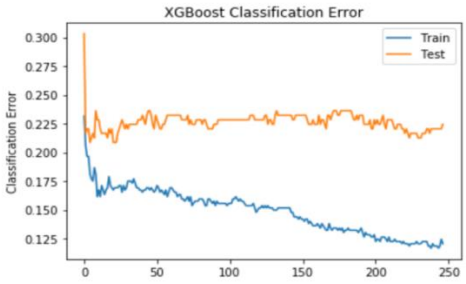
XGBoost we knew the buying product style feature f importance value for each node n for to determine the label y. Along with it we did stacking of the model of XGBoost with the accuracy of 89%. Suppose customer spends more time is one page that page contains the list of products that the customer uses, we can take the customer id as features in which it might be useful information such as ip address, it tells from where the customer is ordering the products, from which place the customer is ordering the products, how many quantities the customer is buying the products. We took the frequency of such information of the customer , it helps business to keep track of the information no the customer and the budgeting factor of the customer. Based on the customer budget the cookies information which we took as the data of the customer especially a cookies information then tracked the data, usually it tells what kind of products the customer likes and dislikes. We took the maximum the another kind of information such as if the customer spends most of the time in

that particular product page, there might be highly likely that customer might like , but he might not purchase that product ,he might look for other factors like costing factor , needs other features for the products , we took the maximum duration of the time which customer spent in that page.

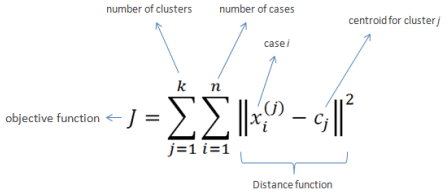
$$f(x) \approx f(a) + f'(a)(x-a) + \frac{1}{2}f''(a)(x-a)^2$$

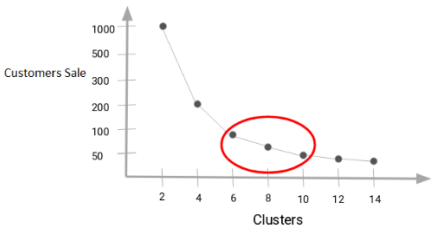
$$\mathcal{L}^{(t)} \simeq \sum_{i=1}^{n}[l(y_i, \hat{y}^{(t-1)}) + g_i f_t(\mathbf{x}_i) + \frac{1}{2}h_i f_t^2(\mathbf{x}_i)] + \Omega(f_t)$$

XGBoost



By clustering and finding the right centroid without the outlier and noise points and reducing it , we did padded sequence whenever needed. Using the MLP that is the deep learning technique by putting the value of weights into the layer of value x at constant b. As the customer increases, the number of orders increases with respect to the number of orders is placed in the system. Internal system became more cumbersome to manage the orders which is given by the customers. We used BFS algorithm internally so that it speeds up and saved lot of time. It increased the productivity at the admin level.

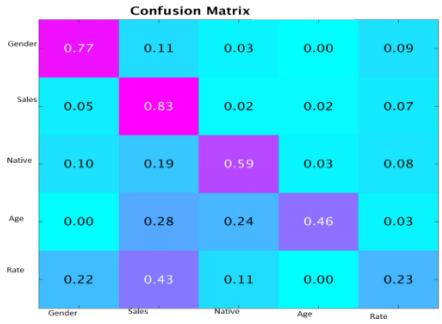




With clustering the data with k number of cluster with n number of points with x values for the c centroid , changing the hyper parameter values for centroid and with the different cluster along with n number of points. In today's business there are multiple ways to deal with the retail dealers. As competation for the business is very large in today's market. Small business has to help the large business owner. This model helps in making the business profitable. Small retail business owner has some of the customers , if they don't find products which the customer demands, they re direct to large shopping mart. This makes the business win win situation for both of them. As large demand of product comes to the large shopping mart, it is impossible to bifericate from which store the demands are coming. We used clustering mechanism , that clusters the data from which shopping mart the data is coming. This makes very clear from which sort of product the small retail customers are demanding at what quantity. So that then system makes it clear then sorting it manually. This made a huge difference in making the biferication part in the admin panel. So this reduces most of human manual work significally less. So making the decision as part computerised with approximately 8 clusters for 200 people found significally less. People working in admin section have less work and people in admin section can be utilized efficently in other department. As we did statistics part as we analysed 13% increase in productivity in the admin section.

$$lrd_k(o) = \frac{\|N_k(o)\|}{\sum_{o' \in N_k(o)} reachdist_k(o, o')}$$

Local reachability distance

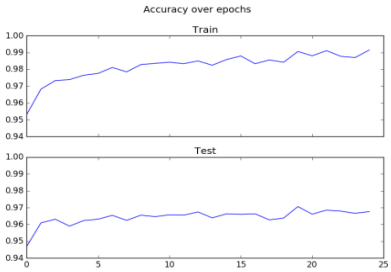


Matrix factorization helped in recommending the products in the store for the classified customers. Using classification neural network we tried to classify use this, before that we did word embedding using keras. Now recommender system helps to get the recommenders product based on the above classification and kd tree technique. As load is increased due to many features such as customer details names for instance names have different alias name. So by this load increases in the system. This is not the only feature. As there are many feature we tend to reduce the dimensions using matrix factorization. As we have used load balancer at server level, at machine learning level also this tends to maintain low level component as hardware is considered. So minimum hardware is required to sort the things out. We take such samples of data from dimension reduction and used in model building, this became easy for preprocessing. Removing stopwords

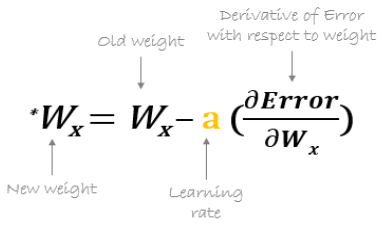
and un wanted words are not considered. When we were implementing these stuff there was no upgrade of the system was needed apart from the development and testing system.

$$y = \varphi(\sum_{i=1}^{n} w_i x_i + b) = \varphi(\mathbf{w}^T \mathbf{x} + b)$$

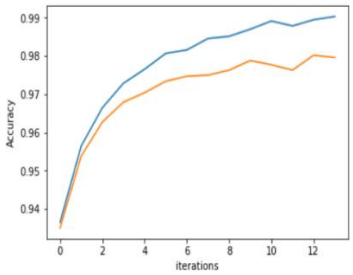
MLP



As error was more which doing the LSTM task, we back propagated using weights along with the learning rate. To increase the accuracy tried with different optimizers, learning rate 0.01 and used few drop out layers. Used sigmoid activation function for dense layer with SGD optimizers with momentum update with learning rate with the accuracy as the number of epoch increases accuracy, increases as error reduces with the accuracy 99%. As there new products have been coming up in the market, names most of the customers would not know it. It is very difficult to describe such products, so we came up with an idea of taking the picture of the product and sending the product details and number of quantity to the retail stores. For first initial days we took the help of whatsapp and email through communication between store executive and customer. So this increased 2% of the sales in the store. But one thing still lagging was speed because executive has to open the email or the whatsapp then read the description of the product then search for the product and for searching for the product the executive has to go to other softwares then search for product it is time consuming although sales has been increased but still the not to the optimal level. We integrated this feature to CRM application so by integrating this feature this becomes easy, executive need not come back and forth now sales increases to 4% in just 15 days of span. To increase the productivity technically we again came across deep learning algorithm which recognize the image. In high level once the customer sends the photo of the product for instance if we take the Parle-G product, by seeing the child's face of words written on it. For that we used image segmentation, where we segmented the image one by one and did elastic search through cloud computing. Now we tried to get the image name and that we segmented the image and did classification based on the product image which we had in our database. Then we did feature analysis some feature which is non important we eliminated using highest value of eigen given. This reduction gave the confidence and performed TSNE above this so that no data gets eliminated along with the variance is not lost. This became the most effective search, as we used SQL search with the query optimization. Got the accuracy of 94% making this into the system with the integration of CRM and ERP application with the sales increase of 8% in 7 days of time.



BackPropogation



Train Accuracy: 0.99
Test Accuracy:0.98

FUNDING

The researcher did not have any funding for this work from any government, commercial or any non –profit organization.

CONCLUSION AND IMPROVEMENT

For performance model we got accuracy using confusion matrix which consists of true positive rate, false positive rate, false negative rate, true negative rate along with Precision and recall with good F1 score 0.95. Doing all the above experiments we came to know that machine learning algorithms like KNN ,Logistic regression, Support vector machine, decision tree, XGBoost and other deep learning techniques like MLP classification technique got 98% accuracy. So choosing MLP technique makes more suitable for the existing data. After that we did the clustering so that all the common language related data gets clustered together using KD tree grouped in terms of segments , so that we can segment the customers , the product which the customer likes or dislikes making these kind of decision helps the management for growth of sales. Bought these specific products which are required for the regional based customers. Once we got the data and to predict the future sale of the store based on the present machine learning learned features, we implemented LSTM for predicting the next month sales. By taking the last few previous data by applying above techniques sales has been improved 20%. At the end of the day delivery is needed to the customer, some of the shopping mall needed box kind of set up. This set up keeps the box shuffling. For instance if we consider the group of boxes which we are ordering from online store. The set of boxes gets ready from the box store, offcourse it needs some IOT device connection. Once the

order keeps coming from the internet store, this comes in the form of queue. This queue structure which is connected to LAN in the store again which is connected to IOT device. That IOT device gets activated and restructure accordingly. Once the order comes in small size the small box gets activated and the order comes in big size and big order gets activated. But dis advantage is to set up the system lot of production cost is involved. But once production is done, we have to reuse the system after certain period of time it will become cost effective. Data base administrator after collecting the data of the customers, not only getting the basic information of the customers, but also getting insights of the customers. This becomes the gold mine of data for the store. Machine learning and deep learning engineers can analyse the data perform modeling which can improve in the process of the system which is directly linked to sale increment.

REFERENCES

1. Alber, Maximilian, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T. Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans. "iNNvestigate neural networks!." J. Mach. Learn. Res. 20, no. 93: 1-8.
2. Alberto, Túlio C, Johannes V Lochter, and Tiago A Almeida. "Tubespam: comment spam filtering on YouTube." In Machine Learning and Applications (Icmla), Ieee 14th International Conference on, 138–43. IEEE.
3. Alvarez-Melis, David, and Tommi S. Jaakkola. "On the robustness of interpretability methods." arXiv preprint arXiv:1806.08049.
4. Ancona, Marco, et al. "Towards better understanding of gradient-based attribution methods for deep neural networks." arXiv preprint arXiv:1711.06104.
5. Apley, Daniel W., and Jingyu Zhu. "Visualizing the effects of predictor variables in black box supervised learning models." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82.4: 1059-1086.
6. Biggio, Battista, and Fabio Roli. "Wild Patterns: Ten years after the rise of adversarial machine learning." Pattern Recognition 84 : 317-331.
7. Borgelt, C. "An implementation of the FP-growth algorithm." Proceedings of the 1st International Workshop on Open Source Data Mining Frequent Pattern Mining Implementations - OSDM '05, 1–5. http://doi.org/10.1145/1133905.1133907 .
8. Cook, R. Dennis. "Detection of influential observation in linear regression." Technometrics 19.1 : 15-18.
9. Dandl, Susanne, Christoph Molnar, Martin Binder, Bernd Bischl. "Multi-objective counterfactual explanations". In: Bäck T. et al. (eds) Parallel Problem Solving from Nature – PPSN XVI. PPSN 2020. Lecture Notes in Computer Science, vol 12269. Springer, Cham .
10. Deb, Kalyanmoy, Amrit Pratap, Sameer Agarwal and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," in IEEE Transactions on Evolutionary Computation, vol. 6, no. 2, pp. 182-197.
11. Doshi-Velez, Finale, and Been Kim. "Towards a rigorous science of interpretable machine learning," no. Ml: 1–13. http://arxiv.org/abs/1702.08608.
12. Emilie Kaufmann and Shivaram Kalyanakrishnan. "Information complexity in bandit subset selection". Proceedings of Machine Learning Research .
13. Grömping, Ulrike. "Model-Agnostic Effects Plots for Interpreting Machine Learning Models." Reports in Mathematics, Physics and Chemistry: Department II, Beuth University of Applied Sciences Berlin. Report 1/2020
14. Heider, Fritz, and Marianne Simmel. "An experimental study of apparent behavior." The American Journal of Psychology 57 (2). JSTOR: 243–59. .
15. Holte, Robert C. "Very simple classification rules perform well on most commonly used datasets." Machine learning 11.1 : 63-90.
16. Hooker, Giles. "Discovering additive structure in black box functions." Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining.
17. Hooker, Giles. "Generalized functional anova diagnostics for high-dimensional functions of dependent variables." Journal of Computational and Graphical Statistics 16.3: 709-732.
18. Inglis, Alan, Andrew Parnell, and Catherine Hurley. "Visualizing Variable Importance and Variable Interaction Effects in Machine Learning Models." arXiv preprint arXiv:2108.04310.
19. Janzing, Dominik, Lenon Minorics, and Patrick Blöbaum. "Feature relevance quantification in explainable AI: A causal problem." International Conference on Artificial Intelligence and Statistics. PMLR.