Comp1804 Report: Predicting The Topic Of Customers’ Banking Questions

Student Id:001293509

Date:12 April 2023

TABLE OF CONTENT

Comp1804 Report:

Predicting the topic of customers' banking questions

Executive Summary:

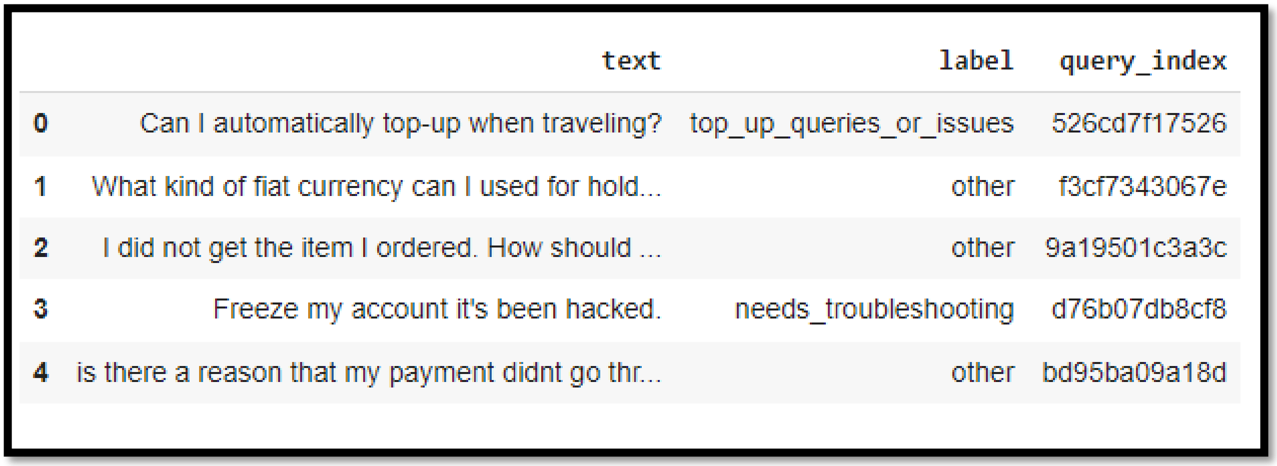
Machine learning is a set of tools and techniques for predicting or categorizing specific occurrences based on known relationships between variables in a given data set (Aderghal et al., 2017). Casanueva et al. first released this dataset, which was obtained from the Moodle link.

Banking is a common empirical and methodological issue for the empirical and methodological study that employs operational research (OR) and artificial intelligence (AI) approaches. This paper uses artificial intelligence and neural networks to present a thorough and systematic way of classifying consumer questions from the banking industry. The research will assist banks in improving customer care through the deployment of bots or automated agents. The analysis concluded that while neural networks outperformed the classic machine learning model, logistic regression also performed well in the banking industry. The analysis concluded that while neural networks outperformed the classic machine learning model, logistic regression also performed well in the banking industry.

Exploratory Data Analysis

Data Description:

The data set was obtained from the Moodle link and is an adaption of the original data set created by Casanueva et al. The following fields make up the data. There are three columns and 14,127 rows in the data set. The primary sample segment of the data set is shown below.



	text	label	query_index
0	Can I automatically top-up when traveling?	top_up_queries_or_issues	526cd7f17526
1	What kind of fiat currency can I used for hold...	other	f3cf7343067e
2	I did not get the item I ordered. How should ...	other	9a19501c3a3c
3	Freeze my account it's been hacked.	needs_troubleshooting	d76b07db8cf8
4	is there a reason that my payment didnt go thr...	other	bd95ba09a18d

Figure 1 Data sample; source: self

The description of the dataset can be viewed using the describe command. The view of describe command shows count, Unique, Top and freq as below.

Semantics

Text – Refers to the text entered by the customer as their question

Label-Refers to the label given to classify the query

Query_index- Refers to the index of query

Dimensionality reduction technique

Principal Component Analysis (PCA) is the most widely used approach for analyzing high-dimensional data. There are several theories on how PCA lowers dimensionality (Chen and Guestrin, 2016). We'll start with the geometric interpretation, where this operation might be understood as a rotation of the data's initial dimensions. The original data space is rotated via PCA such that the axes of the new coordinate system point to the directions of the highest variance in the data. Principal components (PC) are the axes or new variables that are arranged by variance: The first component, PC 1, reflects the data's direction of highest variation. The second component's direction, PC 2, reflects the biggest of the remaining variances orthogonal to the first component. This may be simply expanded to acquire the appropriate amount of elements which span a component space with the correct level of variation.

Data Preprocessing

The first step in data pre-processing is to check whether there are any missing rows or values in the given dataset. To check for null values python supports the command isna(). The isna method returns true value for null columns. The next step is to sum all the null values to know the total number of null values. From the output, it is clear that column ‘label’ has missing values in it.

Data Cleaning

Data cleaning involves various steps, the present dataset has missing values, unrelated data and duplicate values.

Dropping unrelated Data

The data has an unrelated symbol hash in the text column. The text column with hash symbols must be dropped for which we will use the drop method of Python.

Capitalization

The next step is to capitalize all starting characters of sentences in the text column. The method used for capitalization is lambda.capitalize.

Handling missing values

Simpleimputer is used to replace missing data with basic techniques such as mean, median, or most often occurring values. In the present dataset, the technique is to fill empty 'label' variables with the most common occurring values.

Removing Duplicate values

The duplicate values can be removed using methods such as dropna and drop_duplicates. For checking duplicate values, we have also renamed the ‘label’ column data.

Encoding Categorical data

For understanding and learning, machine learning models require numerical data. This means that if the data contains categorical data, it must be numerically coded before the model is fitted and evaluated (Here's Everything You Need to Know About Categorical Data Coding, 2020).

Ordinal coding and one-hot coding are the two most common approaches. In the present encoding, an object is constructed in the tag-encoder class and applied to the category values in the 'tag' column. The first transformation is used in the encoding technique, which is beneficial for converting all data at once.

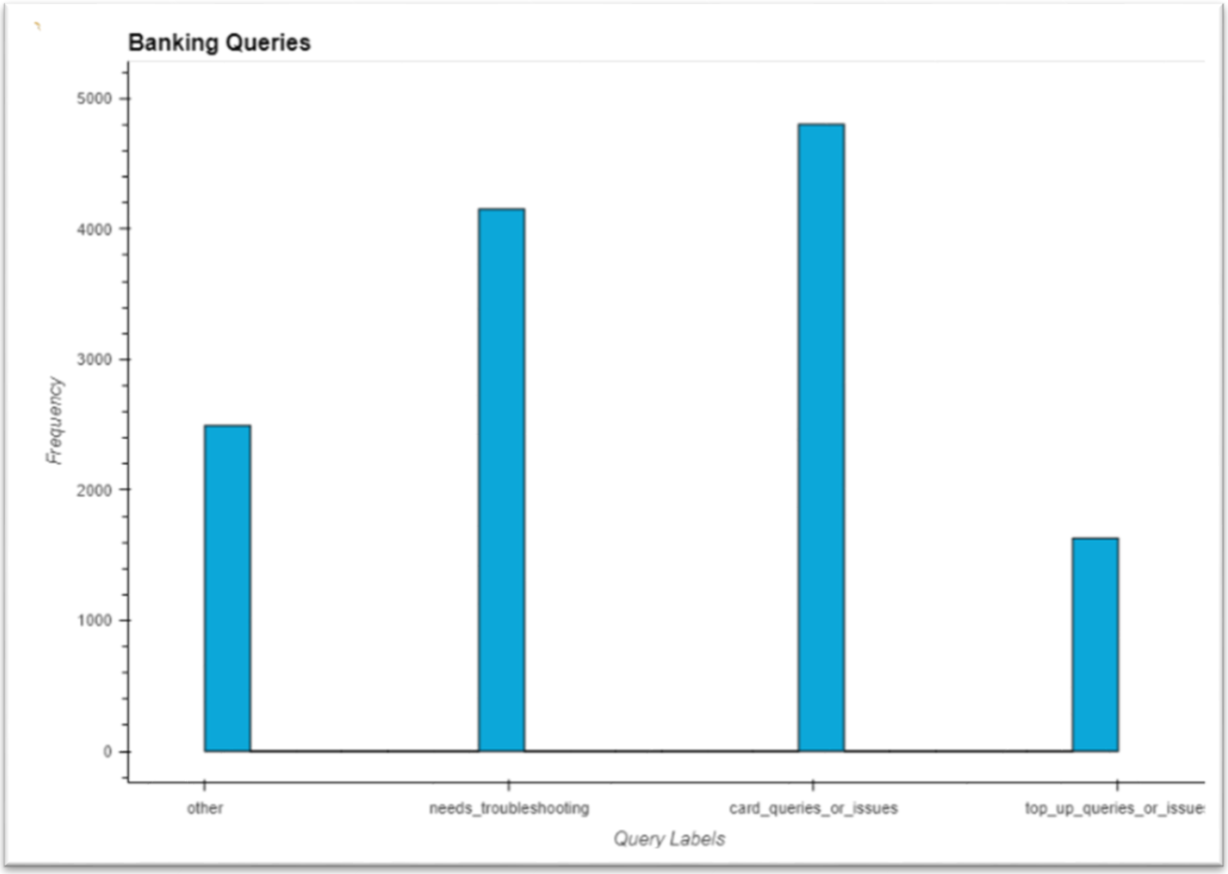


Figure 2 encoding; source: self

Text Pre-processing

The next step after converting categorical data is to pre-process the text which covers removing commas, punctuations and lemmatization. For comma, we will use the removing punctuation function which when called with the text field will pre-process the text and make it punctuation free.

Lemmatization is the process of combining several versions of a word in order to treat the word as a single part or component. We will lamitise stem the words in the label field because there are numerous terms that can be grouped together.

Data splitting

At this point, we will split the data into test and training partitions. The reason is that training creates the data to build the model, and we need an independent test not used for training to test the model's performance. The data has been divided into two parts with 35% of data as testing data and 65% as training data.

Methodology

Our system employs a single-data, multi-model approach, which implies that we investigate the properties of client inquiries using several machine-learning algorithms. A well-studied CNN ResNet was used in the neural networks to exploit the tag category and grasp the issue at hand. Under classic machine learning methodologies, logistic regression, Random Forest, and Naive Bayes were used in the second direction.

Banks, as financial institutions that have direct contact with people, require a deeper understanding of their clients as their most precious asset in order to improve their market share and maximize the utilization of bank resources in the delivery of suitable services. While several factors impact the adoption of these new financial channels, the rise of electronic channels has exacerbated this demand. Banks can enhance customer satisfaction by utilizing machine learning. This research examines utilizing artificial intelligence to forecast the topic of client enquiries in order to assist banks in categorizing inquiries and treating them appropriately (Li et al., 2021).

Machine learning implementation using traditional machine learning

Conventional machine learning techniques (excluding deep learning) encompass a wide range of methodologies, including supervised, semi-supervised, and unsupervised modelling techniques. They are frequently used for data enhancement and dimensionality reduction (Paterakis et al., 2017).

Logistic regression, Random Forest classifier, and Nave Bayes are the methods employed in this paper. Each algorithm's performance measures include accuracy, precision, f1 recall score, and confusion matrix plot. The prediction summary is represented as a matrix in a confusion matrix. The number of right and erroneous guesses for each class is displayed. It aids in comprehending courses that are perplexed by the model as another class.

Naive Bayes Classifier

Bayes' theorem underpins the naive Bayes classifier. It is assumed that each feature is a distinct provided class. Several practical applications, including text categorization, medical diagnostics, and system performance monitoring, have demonstrated the effectiveness of Naive Bayes (Domingos & Pazzani, 1997). Aspects of Naive Bayes' limitations are well known: for binary data, it can only train linear discriminant functions and is hence always poor for separable nonlinear notions. Despite its shortcomings, Naive Bayes has proven to be the best method for several significant classes of ideas with a high degree of feature reliance, such as disjunctive and conjunctive concepts (Domingos & Pazzani, 1997).

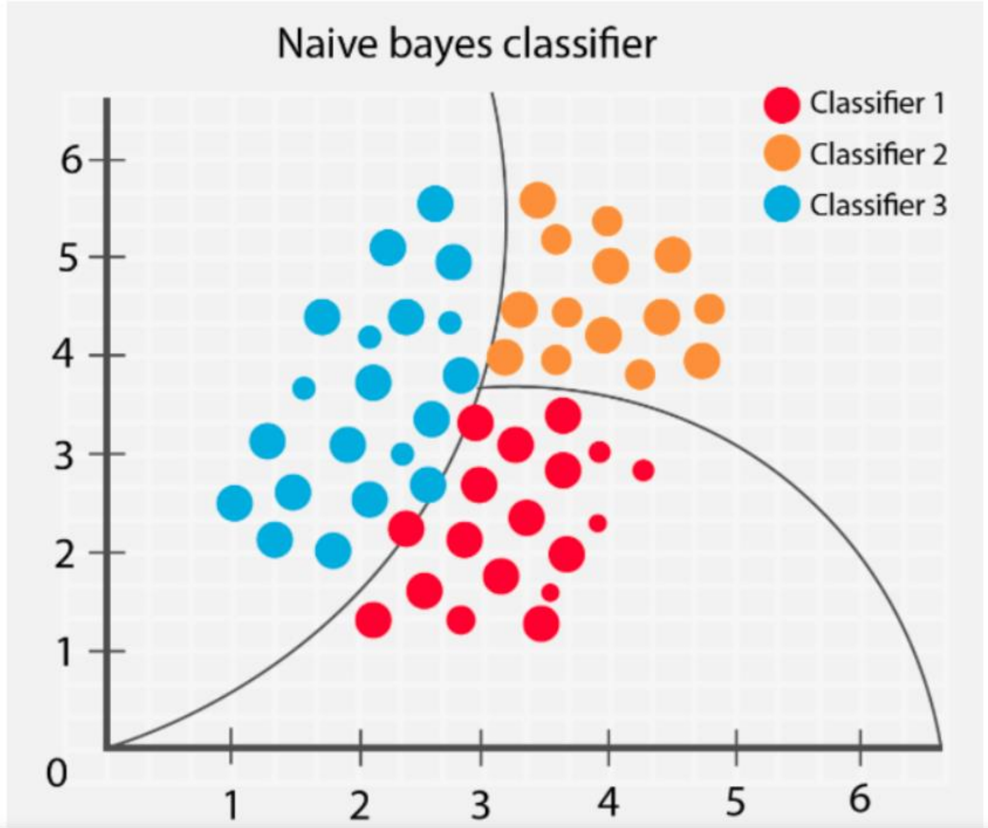


Figure 3 Naive Bayes Classifier; Source: Analyticsvidhya.com, 2023

Because the data set has a text column and the relationships cannot be determined, Naive Bayes was employed in this notion. To compare the performance of Naive Bayes with various algorithms, precision, precision, recall, and f1 score were determined. Naive Bayes has an accuracy of 82%.

Accuracy

The percentage of correctly categorized data instances over all data instances is known as accuracy. The accuracy by Naïve Bayes is 82%

$$Accuracy = \frac{TN + TP}{TN + FP + TP + FN}$$

Figure 4Accuracy formula, Source: Google images

Precision

The precision of an effective classifier should ideally be high with 1 as the value. The precision becomes one only when the numerator and denominator are identical, or when TP = TP + FP, which also means that FP is zero. With increasing FP, the denominator value exceeds the numerator value, but the precision value decreases.

$$\text{Precision} = \frac{\mathit{True\ Positive}}{\mathit{True\ Positive} + \mathit{False\ Positive}}$$

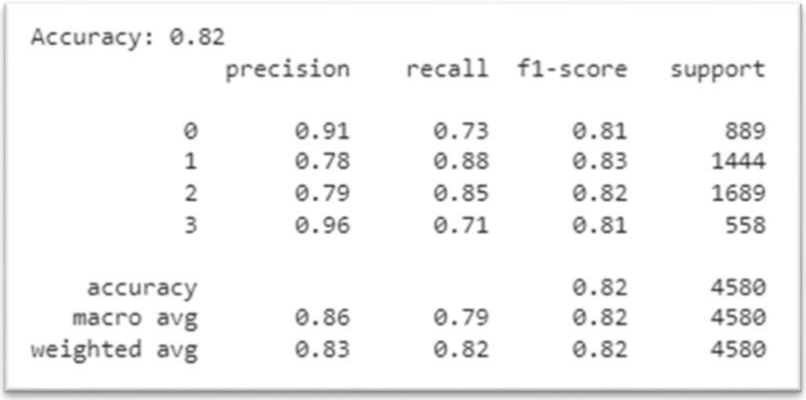
Figure 5Precision formula, Source: Google images

Recall

The recall is also known as the sensitivity or true positive rate. The optimal recall for a capable classifier is one (high). The memory becomes one only when the numerator and denominator are the same, that is, TP = TP + FN; this also implies that FN is zero. When FN grows, the denominator's value exceeds the numerator's, and the payback value declines, leading in a weak AI model.

$$\text{Recall} = \frac{\mathit{True\ Positive}}{\mathit{True\ Positive} + \mathit{False\ Negative}}$$

Figure 6Recall formula; Source: Google images



```
Accuracy: 0.82
              precision    recall  f1-score   support

           0       0.91      0.73      0.81       889
           1       0.78      0.88      0.83      1444
           2       0.79      0.85      0.82      1689
           3       0.96      0.71      0.81       558

    accuracy                           0.82      4580
   macro avg       0.86      0.79      0.82      4580
weighted avg       0.83      0.82      0.82      4580
```

F1 score

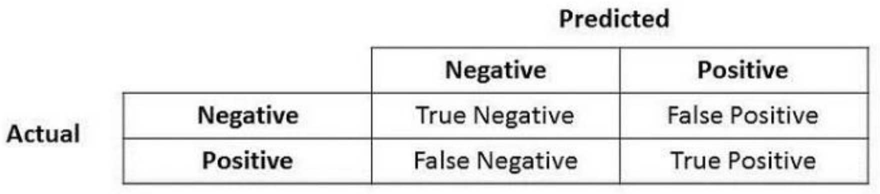
To have a strong classifier, accuracy and recall must be one, which means FP and FN must be zero. As a result, we need a statistic that takes accuracy and recall into account. The F1 score is a metric that takes into account both accuracy and recall.

$$F1 = 2 \times \frac{Precision * Recall}{Precision + Recall}$$

Figure 7 F1 score formula, Sciencedirect.com

Confusion Matrix

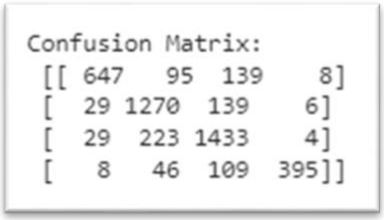
The prediction summary is represented as a matrix in a confusion matrix. It shows the number of correct and incorrect guesses for each class. It helps to understand the classes that are bewildered by the model as another class.



		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

Figure 8confusion matrix, source: science direct

Confusion matrix of Naïve Bayes is as below



```
Confusion Matrix:
 [[ 647   95  139    8]
 [   29 1270  139    6]
 [   29  223 1433    4]
 [    8   46  109  395]]
```

Random forest Classifier

A random forest is a meta estimator that uses the average to increase prediction accuracy and control for overfitting by fitting a group of decision tree classifiers to distinct subsamples of a data set. max samples (Scikit-learn, 2018) governs subsample size. It may also be thought of as a supervised algorithm with numerous decision trees.

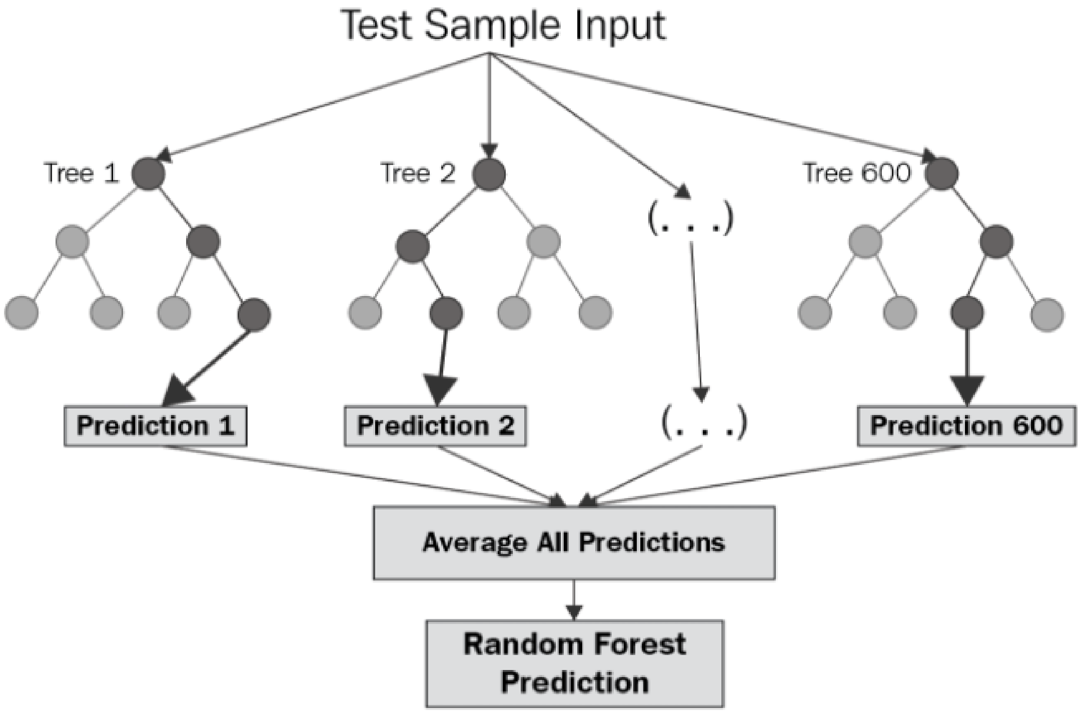
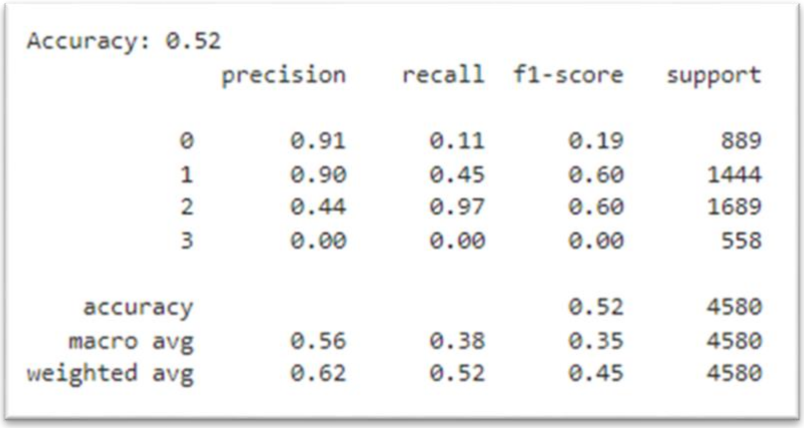


Figure 9 Random Forest, Source: Bakshi, 2020

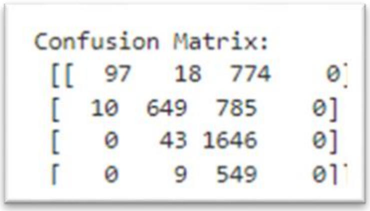
The Random Forest Classifier's performance was validated by measuring accuracy, precision, recall, and f1 score. The Random Forest Classifier has an accuracy score of 52%, which is much lower than Naive Bayes.

Accuracy, Precision, Recall and f1 score table



```
Accuracy: 0.52
              precision    recall  f1-score   support

           0       0.91      0.11      0.19       889
           1       0.90      0.45      0.60      1444
           2       0.44      0.97      0.60      1689
           3       0.00      0.00      0.00       558

    accuracy                           0.52      4580
   macro avg       0.56      0.38      0.35      4580
weighted avg       0.62      0.52      0.45      4580
```

<u>Confusion matrix of random forest</u>



```
Confusion Matrix:
 [[  97   18  774    0]
 [  10  649  785    0]
 [   0   43 1646    0]
 [   0    9  549    0]]
```

Logistic Regression

Logistic regression is a statistical analysis technique in which the concept of individual items is compared to naive bayes. Consider and train on one or more independent variables that are accountable for the outcome (Scikit-learn.org, 2014).

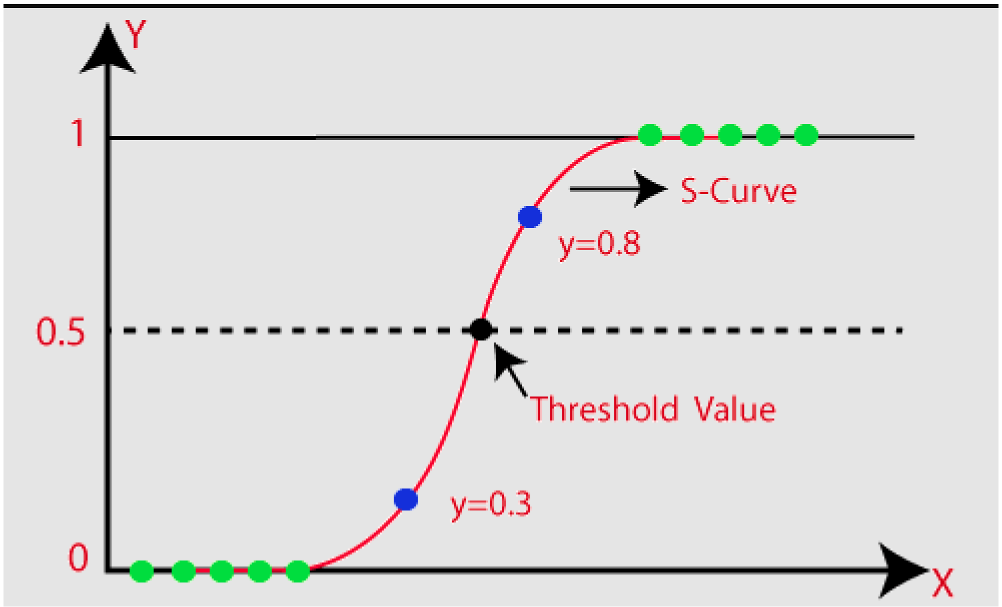
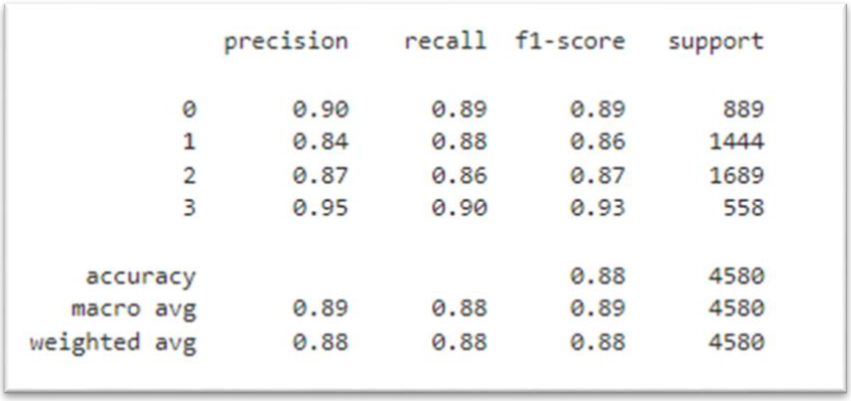


Figure 10 Logistic regression, www.javatpoint.com

Officially, binary logistic regression has a single binary dependent variable coded by an indicator variable, the two values of which are indicated by "0" and "1," whereas the independent variables might be binary or continuous. The chance of a value indicated by '1' varying from 0 (some value '0') to 1 (some value '1').

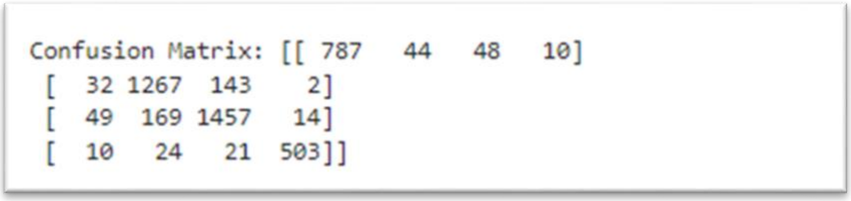
Accuracy, Precision, Recall and f1 score table



```
              precision    recall  f1-score   support

           0       0.90      0.89      0.89       889
           1       0.84      0.88      0.86      1444
           2       0.87      0.86      0.87      1689
           3       0.95      0.90      0.93       558

    accuracy                           0.88      4580
   macro avg       0.89      0.88      0.89      4580
weighted avg       0.88      0.88      0.88      4580
```

Confusion Matrix

The confusion matrix of Logistic Regression is as below:



```
Confusion Matrix: [[ 787   44   48   10]
 [  32 1267  143    2]
 [  49  169 1457   14]
 [  10   24   21  503]]
```

Logistic regression has the greatest accuracy of the three methods used for classical machine learning on the supplied dataset, at 88 per cent.

Machine learning implementation using Neural networks

Artificial neural networks (ANNs), sometimes known as "neural networks" (NNs), are computational models that behave similarly to the human brain, learning from inputs and producing output. An NN is composed of numerous layers. The first and last layers of a conventional feedforward design correspond to input and output, respectively, while the intermediate (hidden) levels have many neurons, each of which works as a processing unit.

For increasingly difficult problems, several NN architectures are created. Convolutional neural networks (CNN; LeCun, 1989) are a family of deep learning architectures that have been effectively applied to data having grid-like topologies, such as bank-collected phone records, through process automation. A neuron in a CNN is the outcome of several convolution procedures before the final results are formed.

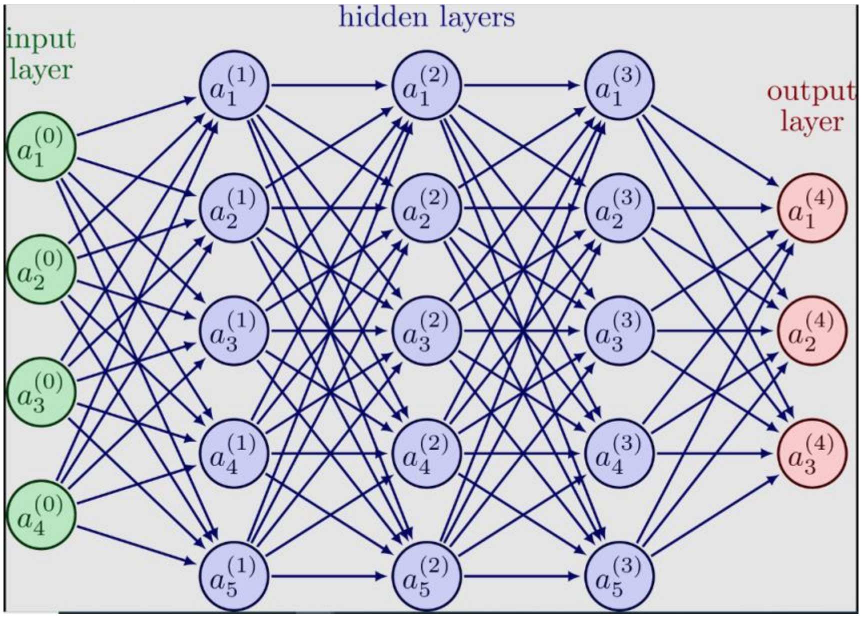


Figure 11 Neural Netword, source: Tikz.net

Several works add to the literature in terms of innovation by constructing NNs based on various sorts of data. Ronnqvist and Sarlin (2017), for example, used a traditional NN to anticipate banking difficulties based on the text inputs that comprise the semantic vectors in the hidden layers. Kriebel and Stitz (2021) used text inputs in a deep learning NN model to predict defaults on P2P loans, while Ladyzynski et al. (2019) used a deep learning system to analyze transactional data for customer relationship management in retail banking, but they discovered that a conjoint approach based on the random forest algorithm performed better. In the current paper, upgraded NNs were used to forecast the topic of client questions in the banking industry. Before applying neural networks to the data collection, additional data preparation is performed. The pipeline procedure is used to transform a sparse matrix into a dense matrix. The model in the training set is summarized here.

Machine learning models are insufficiently sophisticated to determine which hyperparameters will deliver the maximum potential accuracy for a given data set. But, properly adjusting the hyperparameter values can result in a very accurate model, thus during the training phase, the model attempts several hyperparameter combinations and utilizes the optimum combination of hyperparameter values to create predictions. The next step is to define the hyperparameters; the Python code generates the batch size indicating the data points to compute the model parameter update. The model is fitted to the training data set, and its performance is assessed.

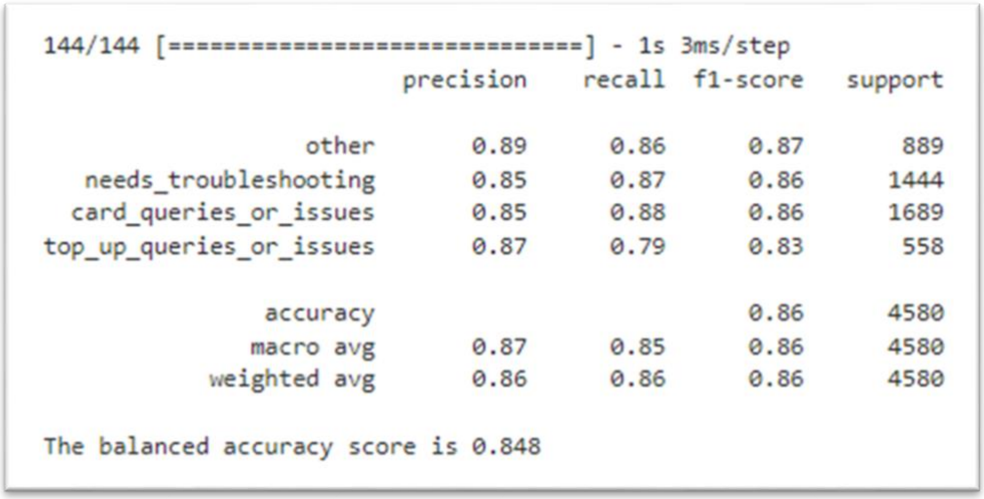
Loss:

The loss value on the training data at the conclusion of each epoch is referred to as "loss." Through training, the optimization process aims to reduce this, thus the lower the better.

Ratio:

The "precision" is defined as the ratio of correct predictions to total predictions in the training data. The greater the value, the better. This is frequently, but not always, inversely proportionate to the loss. Epoch 50/50 has a 97 per cent accuracy rate, which is greater than typical machine learning.

Accuracy, Precision, Recall and f1 score table

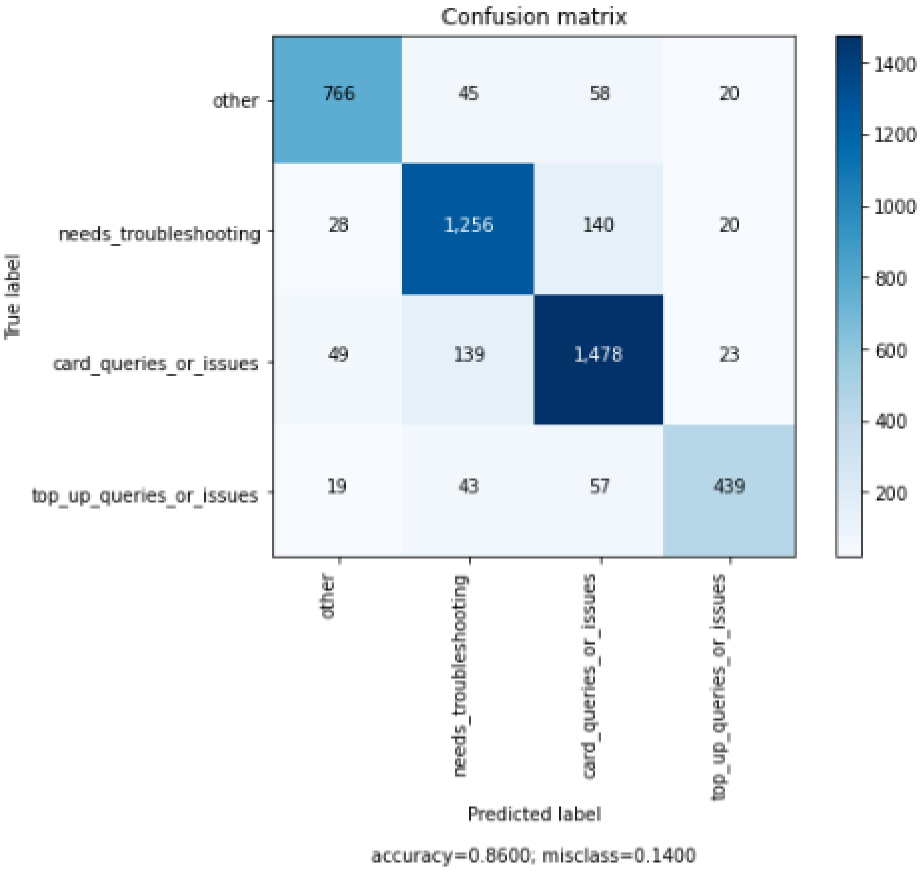


```
144/144 [==============================] - 1s 3ms/step
                          precision    recall  f1-score   support

                   other       0.89      0.86      0.87       889
      needs_troubleshooting       0.85      0.87      0.86      1444
    card_queries_or_issues       0.85      0.88      0.86      1689
  top_up_queries_or_issues       0.87      0.79      0.83       558

                accuracy                           0.86      4580
               macro avg       0.87      0.85      0.86      4580
            weighted avg       0.86      0.86      0.86      4580

The balanced accuracy score is 0.848
```

According to the above results the accuracy score of the model on the test dataset is 86 per cent.

Confusion Matrix

The prediction summary is represented as a matrix in a confusion matrix. The number of right and erroneous guesses for each class is displayed. It aids in comprehending courses that are perplexed by the model as another class.



A comparison between classic machine learning versus neural networks models

The neural network has an accuracy score of 87%, which is lower than typical machine learning models such as Naive Bayes, Logistic Regression, and Random Forest.

The neural network model has a recovery score of 85%, which is lower than the logistic regression model but higher than the Random Forest and Nave Bayes models.

The neural networks have an F1 score of 86%, which is lower than logistic regression but higher than the Random Forest and Nave Bayes models.

The accuracy of the neural networks model is 86%, which is greater than the classic machine learning models Random Forest with 52, Nave Bayes with 82, and Logistic Regression with 88%.

Ethical Considerations

The following are some ethical problems to consider for the provided research study:

- Instead of exhibiting the outcomes of other people's studies by duplicating their code, the study contains its own analysis.
- Just Moodle datasets were utilized for the experiment, with no alterations made to get findings, resulting in honesty throughout the project lifespan.
- To prevent the potential of a copyright claim by authorized people or writers, all data and material to be utilized in the research are correctly referenced.

Ethical OS Toolkit

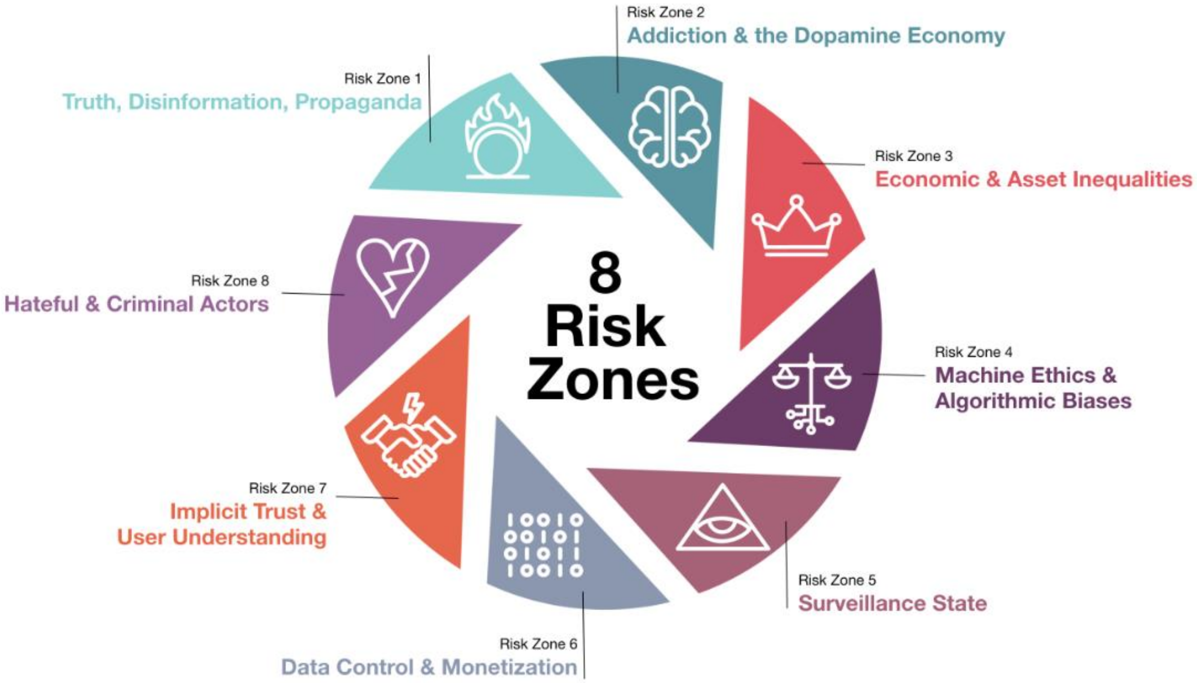

Risk Zone 2
Addiction & the Dopamine Economy
Risk Zone 1
Truth, Disinformation, Propaganda
Risk Zone 3
Economic & Asset Inequalities
8 Risk Zones
Risk Zone 8
Hateful & Criminal Actors
Risk Zone 4
Machine Ethics & Algorithmic Biases
Risk Zone 7
Implicit Trust & User Understanding
Risk Zone 5
Surveillance State
Risk Zone 6
Data Control & Monetization

Figure 12 Ethical Risk zones; wiki images

Risk Zones

Truth, Disinformation, Propaganda

Nobody has the ability to modify the data that has been obtained from Moodle. The dataset contains no information that is directly related to any individual's personal information.

Machine Ethics and Algorithms Biases

The "black box" secrecy surrounding the development of machine learning algorithms is a major source of worry about bias. The for-profit companies that develop these algorithms do not provide the criteria and calculations that go into the formulae. Algorithms are frequently so complex that even engineers and designers having access to the formulae may find it difficult or impossible to predict the outcome and consequences of algorithm findings. As a result, because these algorithms are built by people, they reflect society's views, prejudices, and discriminatory behaviours unwittingly and often unconsciously (Center for Internet and Human Rights, 2017).

To counteract these biases, we can only utilize algorithms that have a track record of generating judgements without prejudice. Nevertheless, there is presently no option available to modify the algorithm to make it bias-free.

Data Control and Monetization

One of the major disadvantages of machine learning and deep learning approaches is the requirement for enormous data sets for development and testing, which are frequently orders of magnitude or bigger than those gathered in the field. most perspectives. Personal and sensitive information may be gathered while the data sets are being collected. Data gathering and usage without authorization pose ethical and legal concerns. The data set utilized in this study was not gathered without consent from individuals, and all personal information about them was concealed.

Implicit Trust and User Understanding

Offering consumers options, and hence control over how they choose to communicate with an agent, is an important first step in addressing their needs and well-being. Transparency about an agent's status as an automated (non-human) agent, for example, and the boundaries of its capabilities is crucial for allowing users to make educated decisions, which leads to user trust. According to recent research, while speaking with an automated agent, people behave and respond differently than when communicating with another human person (Mou and Xu, 2017). When customers are aware that they are engaging with a human or an automated agent, they may make educated decisions about their own conduct, particularly when it comes to information sharing (Gentsch, 2019). The reports assist the automated agent in responding to users by categorizing the inquiry, and no attempt has been made to fool the user through the use of AL models.

Recommendations

After analyzing the data and developing the model, I would recommend neural networks for predictive models, particularly in the banking industry. The neural network enables robots to think like people in order to assist them in making decisions.

- The neural network and logistic regression are the finest machine learning models. Both models performed well and can be used for prediction.
- The final neural network model is suitable for usage in the banking sector due to its high accuracy score and quick calculation.
- Other methods, such as the integration of LSTM and GRU in neural networks and XGBoost, can be used to improve the research.

Conclusion and Retrospective

This article provided a thorough examination of the critical function that traditional AI and NN approaches play in banking research by evaluating user queries and predicting the topic linked to the inquiry. To revise the paper, I would suggest using AWS Studio for data cleaning and integration as it is a fast and reliable tool. And regarding the machine learning model, XGBoost may be more useful when compared to other algorithms in the prediction sector for banks.

References

1. *P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, 29:103–130, 1997.*

2. *Scikit-learn.org. (2014). sklearn.linear_model.LogisticRegression — scikit-learn 0.21.2 documentation. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.*

3. *Scikit-learn (2018). 3.2.4.3.1. sklearn.ensemble.RandomForestClassifier — scikit-learn 0.20.3 documentation. [online] Scikit-learn.org. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html .*

4. *Li, F., Lu, H., Hou, M., Cui, K. and Darbandi, M. (2021). Customer satisfaction with bank services: The role of cloud services, security, e-learning and service quality. Technology in Society, 64, p.101487. doi:https://doi.org/10.1016/j.techsoc.2020.101487.*

5. *Harvard President & Fellows (2017, January 20). Berkman Klein center and MIT media lab to collaborate on the ethics and governance of artificial intelligence - Harvard law today*

6. *Mou, Y., Xu, K.: The media inequality: Comparing the initial human-human and human-ai social interactions. Computers in Human Behavior 72, 432–440 (2017)*

7. *Gentsch, P.: Conversational ai: How (chat) bots will reshape the digital experience. In: AI in Marketing, Sales and Service, pp. 81–125. Springer (2019)*

8. *Aderghal, K., Boissenin, M., Benois-Pineau, J., Catheline, G. and Afdel, K., 2016, December. Classification of sMRI for AD Diagnosis with Convolutional Neuronal Networks: A Pilot 2-D+ Study on ADNI. In MultiMedia Modeling: 23rd International Conference, MMM 2017, Reykjavik, Iceland, January 4-6, 2017, Proceedings, Part I (pp. 690-701). Cham: Springer International Publishing.*

9. *Chen, T. and Guestrin, C., 2016, August. Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).*

10. *Here's All you Need to Know About Encoding Categorical Data (with Python code). (2020). Analytics Vidhya. [online] 14 Aug. Available at: https://www.analyticsvidhya.com/blog/2020/08/types-of-categorical-data-encoding/.*

11. *N. G. Paterakis, E. Mocanu, M. Gibescu, B. Stappers and W. van Alst, "Deep learning versus traditional machine learning methods for aggregated energy demand prediction," 2017 IEEE PES Innovative Smart Grid Technologies*

Conference Europe (ISGT-Europe), Turin, Italy, 2017, pp. 1-6, doi: 10.1109/ISGTEurope.2017.8260289.

12. *LeCun, Y., 1989. Generalization and network design strategies. Connectionism in perspective, 19(143-155), p.18.*
13. *Rönnqvist, S. and Sarlin, P., 2017. Bank distress in the news: Describing events through deep learning. Neurocomputing, 264, pp.57-70.*
14. *Kriebel, J. and Stitz, L., 2022. Credit default prediction from user-generated text in peer-to-peer lending using deep learning. European Journal of Operational Research, 302(1), pp.309-323.*
15. *Ładyżyński, P., Żbikowski, K. and Gawrysiak, P., 2019. Direct marketing campaigns in retail banking with the use of deep learning and random forests. Expert Systems with Applications, 134, pp.28-35.*
16. Casanueva et al. (2020) *Efficient Intent Detection with Dual Sentence Encoders*. Proceedings of the 2nd Workshop on NLP for ConvAI - ACL 2020. Data available at https://github.com/PolyAI-LDN/task-specific-datasets.