Bank Marketing and Classification of Audio

Prof. Dr. Medha Wyawahare, Nisarg Mehta
Department of Electronics and Telecommunication Engineering

***Abstract* —** *About Bank marketing, the main purpose is to predict whether the person would be eligible to take loan or not and if the person subscribes for the term deposits. The project is implemented using Pyspark on Databricks and using Data Science Libraries on Google Colab. Coming to the another project classification of audio, the main purpose is to classify the different audios by downloading 6GBs of data from the internet by using different deep learning algorithms.*

***Keywords* —** *machine learning, deep learning, pyspark, data science, google colab. algorithms*

INTRODUCTION

Coming to my first project, i.e about the bank marketing, the main purpose of this project was to predict that a person would be eligible for taking the loan from the bank or not and if person subscribes for the term deposits by taking a dataset from UCI Machine Learning Repository. This project is implemented using 2 methods. First one is using pyspark libraries on Databricks community edition software and another is using Data Science Libraries on Google Colab software. Also different algorithms performance is tested on the dataset to verify which algorithm is performing better than other algorithms.

Coming to my second project i.e classification of Audio in which there is the presence of sounds of 10 different classes which are air_conditioner, engine_idling, drilling, dog_bark, children_playing, jackhammer, Street_music, siren, car_horn, gun_shot. In total there are 8733 sound files of total 10 classes, so for each class, there are 870 sound files for processing. Here also different deep learning algorithms are applied to check the performance of the models against this huge dataset.

I. Literature Review

Quite a few researchers have been working on similar problems in recent years and their efforts have been very well documented in publications worldwide. A few different approaches have been tried out to counter this issue but given the nature of the problem, a few shortcomings have stood out no matter what method has been put to use.

Bank Marketing -

1. Bank Direct Marketing Analysis of Data Mining Techniques[1]
   This paper introduces analysis and applications of the most important techniques in data mining; multilayer perceptron neural network (MLPNN), tree augmented Naïve Bayes (TAN) known as Bayesian networks, Nominal regression or logistic regression (LR), and Ross Quinlan new decision tree model (C5.0). The objective is to examine the performance of MLPNN, TAN, LR and C5.0 techniques on real-world data of bank deposit subscription. The purpose is increasing the campaign effectiveness by identifying the main characteristics that affect a success (the deposit subscribed by the client) based on MLPNN, TAN, LR and C5.0.

2. Knowledge creation in banking marketing using machine learning techniques [2]
   The aim of the project is to find how to use machine learning techniques for analysis and making predictions using existing dataset in banking marketing. To find how they can be used together in a process of converting raw data to effective decision making knowledge. Building the predictive models will help to

predict whether the client will subscribe for a term deposit.

3. Using Data Mining Techniques for Detecting the Important Features of the Bank Direct Marketing Data [6]
   In this paper, data mining techniques are used to interpret and define the important features to increase the campaign's effectiveness, i.e., if the client subscribes to the term deposit. The bank marketing dataset from the University of California at
   Irvine Machine Learning Repository has been used for the proposed paper. We consider two feature selection methods namely information gain and Chi-square methods to select the important features. The methods are compared using a supervised machine learning algorithm of Naive Bayes. The experimental results show that a reduced set of features improves the classification performance.

Classification of Audio -

4. Content analysis for audio classification and segmentation [4]
   In this paper,there is the study of audio content analysis for classification and segmentation, in which an audio stream is segmented according to audio type or speaker identity. Audio classification is processed in two steps, which makes it suitable for different applications. The first step of the classification is speech and nonspeech discrimination. In this step, a novel algorithm based on K-nearest-neighbor (KNN) and linear spectral pairs-vector quantization (LSP-VQ) is developed. The second step further divides nonspeech class into music, environment sounds, and silence with a rule-based classification scheme. A set of new features such as the noise frame ratio and band periodicity are introduced and discussed in detail.

5. Audio Signal Classification [5]
   This paper presents the background necessary to understand the general research domain of ASC, including signal processing, spectral analysis, psychoacoustics and auditory scene analysis. Also presented are the basic elements of classification systems. Perceptual and physical features are discussed, as well as clustering algorithms and analysis duration. Neural nets and hidden Markov models are discussed as they relate to ASC.

II. Methodology

A) Bank Marketing

Talking about Bank marketing, this project is implemented using 2 methods. First one is using pyspark libraries on Databricks community edition software and another is using Data Science Libraries on Google Colab software. About the first method, i.e about pyspark, PySpark has been released in order to support the collaboration of Apache Spark and Python, it actually is a Python API for Spark. In this, I have applied the Logistic regression, Decision trees and Random forest to compare the algorithms in terms of different parameters like Accuracy, Precision, Recall and many other parameters. Also have done hyperparameter tunings using 5-fold cross validation to evaluate the models corresponding to these algorithms and evaluated different parameters.

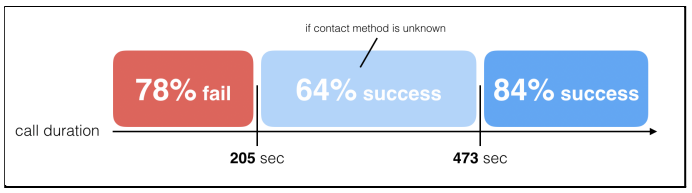

Fig1 : Comparison during call[10]

Talking about the second method, Using Data Science Libraries on Google Colab - In summary, the first cleaning and preprocessing of data is being done. Then I have compared the different parameter performance against the target variable. Then I applied 6 Machine learning Algorithms and compared the train and test accuracy and plotted the ROC Curve for it. The 6 machine learning Algorithms are : Logistic Regression, Random Forest, Support Vector Machine, XGBoost i.e extreme gradient boosting, Stacking Classifier, Voting

Classifier. XGBoost, Stacking Classifier, Voting Classifiers are the advanced algorithms which are applied to the datasets.

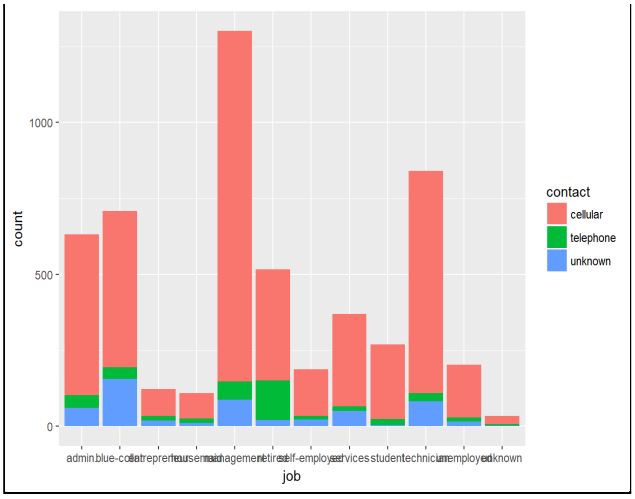

Fig2 : Bar chart for different modes of contact [11]

B) Classification of Audio

Coming to the project of Classification of Audio, in which there is the presence of sounds of 10 different classes. In total, it is 6GBs of data and there are 8733 sound files of total 10 classes, so for each class, there are 870 sound files for processing. Here also different deep learning algorithms are applied to check the performance of the models against these huge datasets.

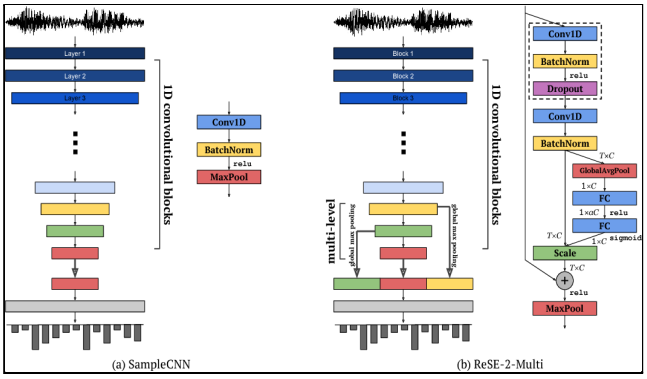

Fig3 : Audio classification using CNN[12]

The main inbuilt library used is librosa and second one is scipy.io. The first step is to analyze the data, i.e it is the data preprocessing step, then I am doing the Exploratory data analysis (EDA) and creating feature extractor on this dataset and finally doing the modeling part and the validation part on this dataset. Taking into consideration the Artificial neural network, I have used the Sequential model, with Dense, Dropout, Activation, Flatten layers into consideration and have used Adam optimizer. The activation function used is 'relu' and dropout used is 0.5.

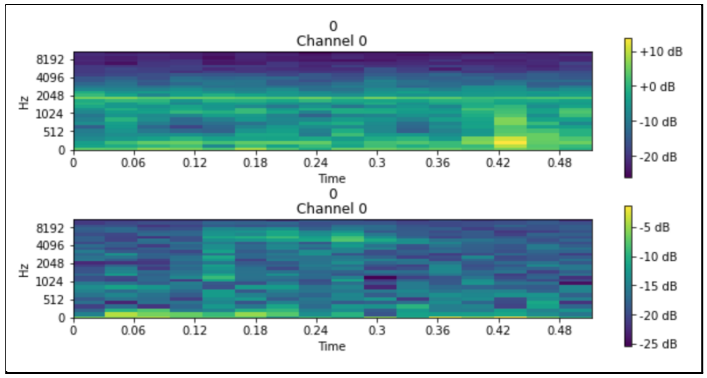

Fig4 : audio classification image[12]

I also did the performance analysis of different algorithms that involve extreme gradient boosting (XGBoost), Convolutional Neural Network(CNN) and random forest. With XGBoost and Random Forest we get the accuracy of 90% and with the CNN, the accuracy is 88% with 64 epochs.

III. Results And Discussions

A) Bank Marketing

Method 1 - Using Pyspark libraries on databricks community platform.

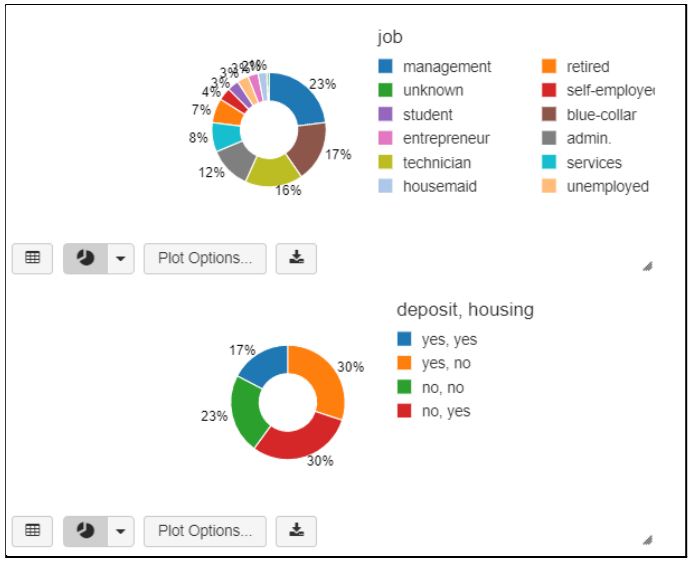

Fig5 : EDA on the data

features	label	probability	predictionLabel
(30,[0,11,13,16,1...	yes	[0.28839605044088...	yes
(30,[0,11,13,16,1...	yes	[0.72994273958118...	no
(30,[0,11,13,16,1...	no	[0.85550652502828...	no
(30,[0,11,13,16,1...	yes	[0.08780852827922...	yes
(30,[0,11,13,16,1...	yes	[0.37922953722179...	yes
(30,[0,11,13,16,1...	yes	[0.57898243636579...	no
(30,[0,11,13,16,1...	yes	[0.04315497658152...	yes
(30,[0,11,13,16,1...	yes	[0.10541328752656...	yes
(30,[0,11,13,16,1...	no	[0.89142071938115...	no
(30,[0,11,13,16,1...	yes	[2.59592672419203...	yes
(30,[0,11,13,16,1...	yes	[0.26751087258647...	yes
(30,[0,11,13,16,1...	yes	[0.03599804825142...	yes
(30,[0,11,13,16,1...	yes	[0.46644715490476...	yes
(30,[0,11,13,16,1...	yes	[0.40381882949893...	yes
(30,[0,11,13,16,1...	yes	[0.47753698199753...	yes
(30,[0,11,13,16,1...	yes	[1.97101641009176...	yes
(30,[0,11,13,16,1...	no	[0.80527956326497...	no
(30,[0,11,13,16,1...	yes	[0.62450972105339...	no

Fig6 : Feature engineering on the data

Table for comparison

Model	Accuracy	Precision	Recall	Area under PR	Area under ROC
Logistic Regression	0.79	0.82	0.80	0.75	0.79
Decision tree	0.78	0.76	0.78	0.73	0.78
Random forest	0.80	0.79	0.80	0.75	0.80

Method 2 - Using Data science libraries on google colab

********** Comparison of different models **********

Model	Test AUC	Test Accuracy
XGBoost	0.853617292665249	0.7851376755501492
Stacking Classifier	0.8637562081932475	0.8793541966161672
Voting Classifier	0.8936204301888323	0.8980426849496849
Logistic Regression	0.8868714117887895	0.7455490434590291
Random Forest	0.7942039265842534	0.8125622028088024
SVM	0.8010343223410121	0.8022780050868075

Fig7 : Comparison of different models

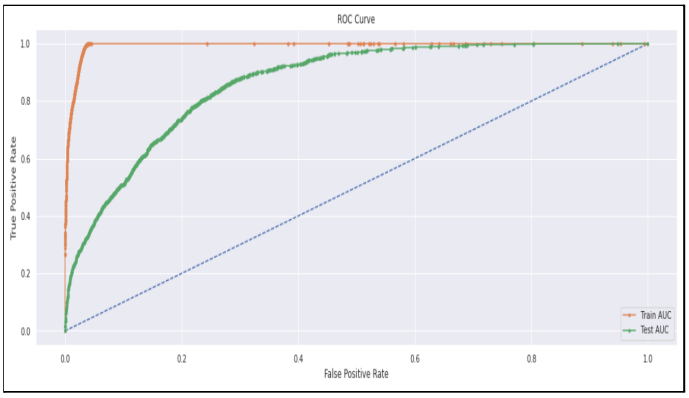

Fig8 : ROC Curve of XGBClassifier

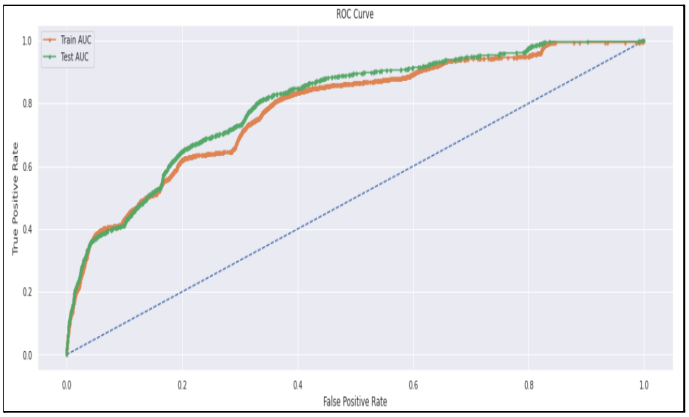

Fig9 : ROC Curve of SVM

B) Classification of Audio

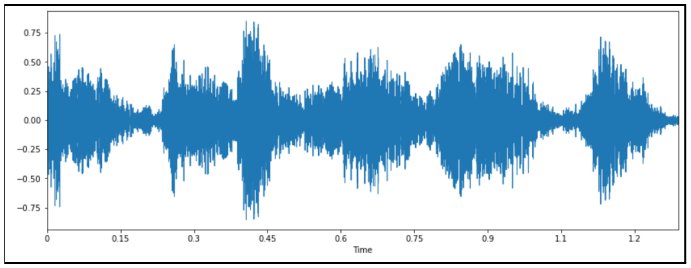

Fig10 : Wave for barking of the dog

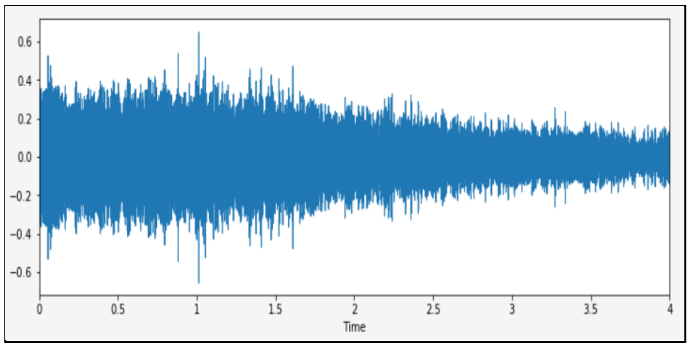

Fig11 : Wave for drilling

```
print(mfccs_scaled_features)
```

```
[  3.3447955   51.35126    -17.376879    10.9488735  -28.23291
   4.4747877  -36.159103    -1.0772134  -34.593834    -7.241792
 -23.64057     -1.8921492  -28.09893    -12.419095   -35.36841
  -8.000556   -21.38167      2.0972068  -17.21334     -6.632617
 -16.367895    -1.0526675  -11.490276    -0.73666435 -12.03093
   0.16736217  -5.2065277   -3.2373405   -4.6389775   -1.7377952
  -0.9288924   -1.2877915   -5.3351254   -1.8893894   -9.446103
   3.217907    -4.5200815   -0.40972534  -0.96236503  -1.3379534 ]
```

```
mfccs_scaled_features=mfccs_scaled_features.reshape(1,-1) # Transformed
print(mfccs_scaled_features)
```

```
[[  3.3447955   51.35126    -17.376879    10.9488735  -28.23291
    4.4747877  -36.159103    -1.0772134  -34.593834    -7.241792
  -23.64057     -1.8921492  -28.09893    -12.419095   -35.36841
   -8.000556   -21.38167      2.0972068  -17.21334     -6.632617
  -16.367895    -1.0526675  -11.490276    -0.73666435 -12.03093
    0.16736217  -5.2065277   -3.2373405   -4.6389775   -1.7377952
   -0.9288924   -1.2877915   -5.3351254   -1.8893894   -9.446103
    3.217907    -4.5200815   -0.40972534  -0.96236503  -1.3379534 ]]
```

```
predicted_label=model.predict_classes(mfccs_scaled_features)
print(predicted_label)
```

```
[7]
```

```
prediction_class = labelencoder.inverse_transform(predicted_label)
prediction_class
```

```
array(['jackhammer'], dtype='<U16')
```

Fig12 : prediction of the sound

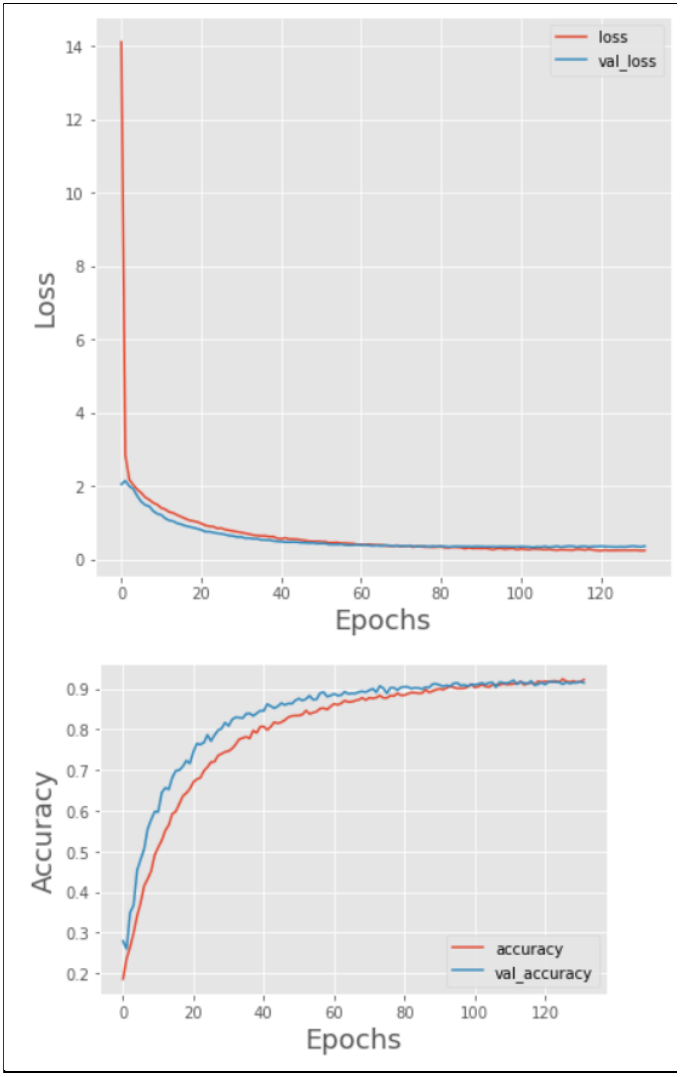

Fig13 : Plot of Loss vs epochs and Accuracy vs epochs

```
Recall: [0.95652174 0.86428571 0.87581699 0.83647799 0.9028777  0.94736842
 0.82474227 0.97704918 0.95131086 0.82450331]
Precision: [0.9625     0.96031746 0.81707317 0.90169492 0.89964158 0.95744681
 0.88888889 0.91975309 0.93040293 0.82178218]

 clasification report:
              precision    recall  f1-score   support

           0       0.96      0.96      0.96       322
           1       0.96      0.86      0.91       140
           2       0.82      0.88      0.85       306
           3       0.90      0.84      0.87       318
           4       0.90      0.90      0.90       278
           5       0.96      0.95      0.95       285
           6       0.89      0.82      0.86        97
           7       0.92      0.98      0.95       305
           8       0.93      0.95      0.94       267
           9       0.82      0.82      0.82       302

    accuracy                           0.90      2620
   macro avg       0.91      0.90      0.90      2620
weighted avg       0.90      0.90      0.90      2620
```

Fig14 : Result for XGBoost classifier

IV. Future Scope

The future scope for bank marketing projects is to make a dashboard and to take live input from the user and to display whether a person is eligible and interested for term deposit or not. And the future scope for the classification of Audio projects is to make the dashboard, where the person can upload the sound file and it would predict the particular sound.

V. Conclusion

For the Bank marketing project, I conclude that with the pyspark using databricks community edition, random forest performed better than logistic regression and decision tree and with the approach of data science libraries on google colab, voting classifier had performed better than rest of the algorithms. For my second project, that is classification of audio, comparing the XGBoost, Convolutional neural network and random forest, I found that XGBoost performed better than the rest of the algorithms.

Acknowledgment

I AM EXTREMELY THANKFUL TO MY PROJECT GUIDE FOR CONTINUOUS SUPPORT, WHO PROVIDED INSIGHT AND EXPERTISE THAT GREATLY ASSISTED ME FOR THIS PROJECT.

References

[1] Bank Direct Marketing Analysis of Data Mining Techniques by Hany A. Elsalamony

[2] Knowledge creation in banking marketing using machine learning techniques by Yana Mihova

[3] Content analysis for audio classification and segmentation by Lie Lu's, Hao Jiang

[4] Audio Signal Classification by David Gerhard and Hong-Jiang Zhang.

[5] Classification of Imbalanced Banking Dataset using Dimensionality Reduction by B. Valarmathi; T. Chellatamilan; Hritik Mittal; Jagrit Jagrit; Shubham Shubham

[6] Visualization and Analysis in Bank Direct Marketing Prediction by Alaa Abu-Srhan, Bara'a Alhammad, Sanaa Al zghoul.

[7] Implementation of the Sound Classification Module on the Platform with Limited Resources by Nives Kaprocki; Nenad Pekez; Jelena Kovačević.

[8] Audio classification using acoustic images for retrieval from multimedia databases by I. Paraskevas; E. Chilton.

[9] Audio classification from time-frequency texture by Slotine, Jean-Jacques E.; Yu, Guoshen.

[10]https://medium.com/@jameschen_78678/which-customers-are-more-likely-to-respond-to-banks-marketing-campaigns-3f00c512268d

[11] https://rpubs.com/pavan721/bank_market

[12]https://towardsdatascience.com/tagged/audio-classification