CAPSTONE PROJECT WITH MACHINE LEARNING

BY SAYANDIP GHOSH & KAUSTAV KONAR

PG CERTIFICATE COURSE IN DATA SCIENCE , AI/ML & DATA ENGINEERING BY E&ICT ACADEMY , IIT ROORKEE, BATCH 2023-'24

PROJECT TITLE

RAIN PREDICTION OF TOMORROW IN AUSTRALIA BY ARTIFICIAL NEURAL NETWORK

Introduction

Artificial Neural Networks (ANNs) are computational models inspired by the human brain. They consist of interconnected nodes organized in layers, and they learn from data through a training process where weights between nodes are adjusted. ANNs excel at tasks like classification, regression, and pattern recognition. Despite their power, challenges remain, such as the need for large datasets and interpretability of results.

OBJECTIVE

To develop and implement an innovative application of Artificial Neural Networks to solve a real-world problem, aiming to demonstrate a deep understanding of ANN's principles, optimize their performance, and contribute valuable insights to the field of machine learning.

Here we use weather forecast related data set over the rainfall in Australia and predict the rainfall of next day depending on the previous day's results of rainfall and weather report by making a Machine Learning model using ANN. Here we see how our model predicts the rainfall report and give the accuracy with the actual data set.

ANN’S APPROACH TO SOLVE THE PROBLEM

In order for a network to perform rainfall prediction, it must understand the input data set. This requires taking the raw data as input that converts the raw information into a complex understanding of the features present within the report. Starting from the network's input layer the data is fed into the network. Each neuron in the input layer passes its signal to neurons in the next layer, called the hidden layer, where computations are performed based on weights and biases analyzed from our trained data set. This process continues through multiple hidden layers until the final output layer produces the network's prediction.

Problem Statement

Predict tomorrow’s rainfall from Australia’s Weather Data using Artificial Neural Network.

What we solve

Here our data set is imbalanced and using this we aim to make a machine learning model by solving this problem which will give more than 80% accuracy in prediction.

PROCEDURE OF MAKING MODEL

- The raw data is set to get pre-processed for handling the missing values , null values , unnecessary columns , categorical variables and feature scaling and dividing the final data into test and train set .
- Model Architecture for an ANN model is constructed using TensorFlow and Python3 in Kaggle Notebook . The architecture includes an input layer, one or more hidden layers, and an output layer. Activation functions and batch normalization technique .
- The ANN model is trained using the preprocessed training dataset. Training process involves forward propagation, backpropagation, and weight optimization using gradient descent algorithms. The model is iteratively trained on the data to minimize the prediction error. The trained model is evaluated using the preprocessed testing dataset. Evaluation metrics such as r2-score , f1 score , confusion matrix , precision , recall are calculated for predicting the Output. Here we see high accuracy cause it hugs the upper left corner of plot indicating high sensitivity and low false positive rate in ROC curve.
- The trained model can be used to make predictions on new, unseen weather - rainfall data. The input data needs to be preprocessed in the same manner as the training data before being fed into the mode

ANN MODEL

```
import keras
from keras.layers import Dense
from keras.models import Sequential

model  = Sequential()
model.add(Dense(256, activation = 'relu', input_shape = (19, )))
model.add(Dense(90, activation = 'sigmoid'))
model.add(Dense(88, activation = 'sigmoid'))
model.add(Dense(45, activation = 'sigmoid'))
model.add(Dense(2, activation = 'sigmoid'))
model.add(Dense(1, activation = 'sigmoid'))
model.compile(optimizer = 'adam', loss = 'binary_crossentropy', metrics = ['ac
curacy'])
model_history = model.fit(X_train, y_train, validation_data = (X_test, y_tes
t),
                          epochs = 100)
```

MODEL SUMMARY

- **Exploring The Dataset & Exploratory Data Analysis**

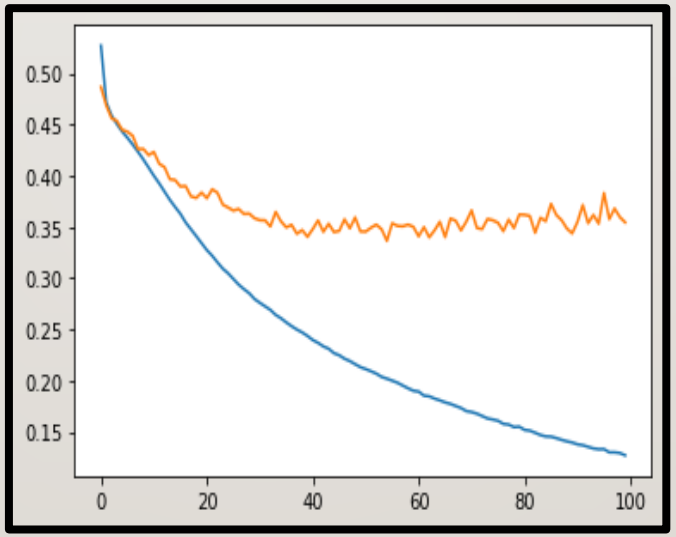
The dataset which we use here for prediction is sourced from Kaggle Dataset Library , providing a weather dataset of Australia's Rainfall . In our dataset there are many independent variables like cloud3pm , cloud9pm , humidity3pm , humidity9am etc which play a significant role in predicting next day's rainfall . Understanding the dataset is important so we can easily create ML model by ANN which will give perfect accuracy.

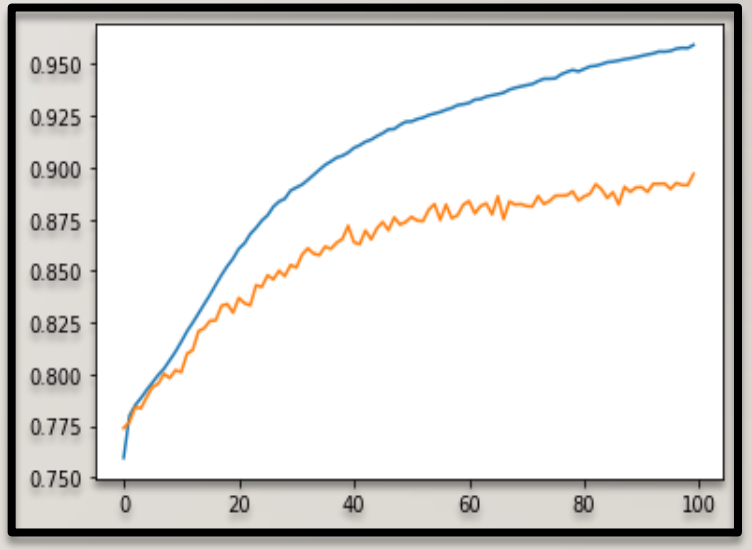
Dataset : https://www.kaggle.com/datasets/arunavakrchakraborty/australia-weather-data/data

MODEL SUMMARY

- Now conclude our prediction observing the train-loss and valid-loss with train and valid accuracy .We also see the r2 score and accuracy-value to know the model's goodness

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 256)               5120
_________________________________________________________________
dense_1 (Dense)              (None, 90)                23130
_________________________________________________________________
dense_2 (Dense)              (None, 88)                8008
_________________________________________________________________
dense_3 (Dense)              (None, 45)                4005
_________________________________________________________________
dense_4 (Dense)              (None, 2)                 92
_________________________________________________________________
dense_5 (Dense)              (None, 1)                 3
=================================================================
Total params: 40,358
Trainable params: 40,358
Non-trainable params: 0
```





MODEL SUMMARY

- Dataset has many null values and some input values of rain tomorrow column is missing that makes our dataset imbalanced and incomplete so we will cover those errors in order to make our model perfectly .
- In our data , by EDA we see that with rain tomorrow , humidity , min-temp ,windspeed , Wind Gust Speed , cloud-level , rain-today and rainfall correlates positively.
- After balancing the dataset , we split the data into train and test part .
- Using standard scale we fit the train and test part . Then make model using sequential and ANN formation by creating hidden layers and using binary cross entropy as activation function with 100 epochs.

RESULTS

Performance Metrics

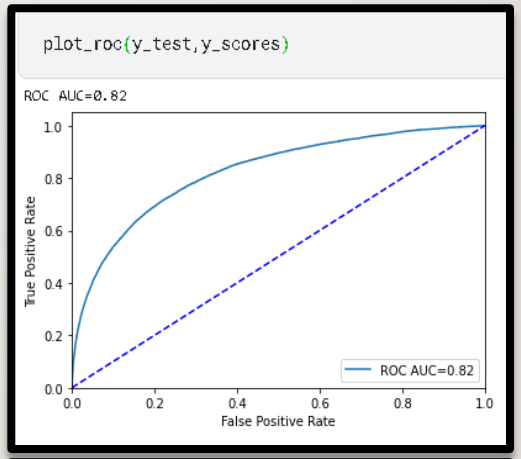
- In our Artificial Neural Network (ANN) project, several performance metrics are commonly used to evaluate the effectiveness of the model in solving our problem. Some of the key performance metrics for ANN projects include :

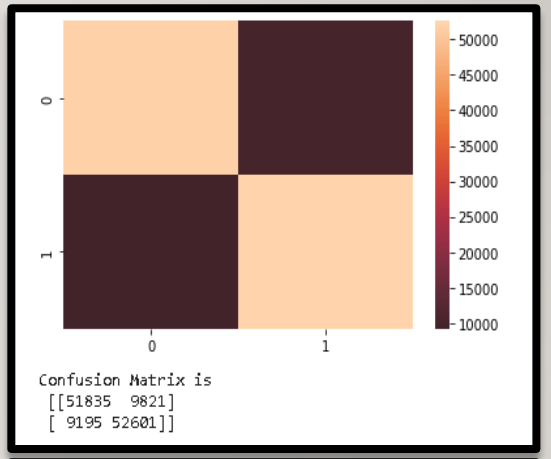
- **Accuracy :** Here we get 89% accuracy measures the proportion of correctly classified instances, out of the total instances present in our data .Our model predicts the correct outcome.

- **R-squared (R2):** R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It's commonly used to evaluate the goodness of fit of a regression model. Here we get R2 score 0.63 which indicates our model is very well trained.

RESULTS

From the logistic Regression Curve we see that ROC curve hugs the upper left corner of the plot, indicating high sensitivity and low false rate ROC AUC value 0.82 and observing other performance metrics we make a random forest classification model with more detailed breakdown by showing True Positive , True Negative , False Positive , False Negative predictions in confusion matrix where we get 84% accuracy , f1 score , precision score and recall score also around 84-85% which denotes our ML model's goodness in the comparison with other performance metrics by using different another classifier model .





RESULTS

Comparison with Existing

- In our model's prediction, the training loss curve is very low compared to the valid loss curve and due to that the accuracy is very high. Which denotes our model's performance is very good.

Novelty

- The novelty in our approach makes it stand out from existing methods. The originality, contributions, and potential impact of our methodology is in a different way of model prediction. Here we use ANN for identifying and highlighting our method's uniqueness.

CONCLUSION

- **By doing this capstone project we learned the process of ANN in ML and how it works over the real time data set. Here we observed that if the input data set is more appropriate and calculative then the output will be more accurate. In ANN model if we increase the no of hidden layers and increase the epoch then our model will be more accurate . Also we use this dataset in different model procedure like SVM, Decision tree etc. Here we also prepare roc curve , confusion matrix in different way . In future we will stick to the above points to get more accuracy in the field of weather forecast.**

Reference & Links

- **We observe several projects in Kaggle library , Cloud-x-Lab Course,YouTube to make our project**
- **Dataset :** https://www.kaggle.com/datasets/arunavakrchakraborty/australia-weather-data/data
- **Python code :** https://github.com/sayandip30882636/sayandip_repo/blob/2fb379986b45054e25aa35ed561a8c111e5b3f0b/Prediction-of-rain-tomorrow-in-australia-by-ANN%20FINAL(1).ipynb