PROJECT REPORT

🏀MARCH MACHINE LEARNING MANIA 2021 NCAAM🏀

TEAM 2

20wh1a6606-J.Akanksha Sharma

20wh1a6632-K.Nikitha Reddy

20wh1a6637-T.Harshitha

20wh1a6638-Sk.Dil Nawaz Farida

20wh1a6646-B.Sreevidya

21wh5a6602-T.Nehasri

INTRODUCTION

Each season there are thousands of NCAA basketball games played between Division I men's teams, culminating in March Madness®, the 68-team national championship that starts in the middle of March.March Machine Learning Mania challenges data scientists to predict winners and losers of the men's 2021 NCAA basketball tournament.

Approach:

The dataset is taken from the NCAA records, which includes all the records of number of teams participating,winning team and their location and many more.

Using those particular records we predict the outcome of the 2021 tournament.

The task is model building and provide a means to score predictions. The real competition is forecasting the 2021 results.

Relevance to real world:

The model would help in predicting results of the 2021 tournament.

TASK DEFINITION

Columns used to build a simple prediction model and submit predictions:

- Team ID's and Team Names
- Tournament seeds since 1984-85 season
- Final scores of all regular season, conference tournament, and NCAA® tournament games since 1984-85 season
- Season-level details including dates and region names.

METHODOLOGY

DATA PREPROCESSING:

- Downloading the dataset
- Link: https://www.kaggle.com/competitions/ncaam-march-mania-2021/data
- Analyzing the dataset and column.
- Delete the unproductive columns

MODELS USED

Different types of python libraries such as pandas, Sklearn, NumPy, matplotlib are used for processing the algorithms.

- LightGBM :

LightGBM is a gradient boosting framework that uses tree based learning algorithms. It is designed to be distributed and efficient with the following advantages:

1. Faster training speed and higher efficiency.
2. Lower memory usage.
3. Better accuracy.
4. Capable of handling large-scale data.

- RANDOM FOREST :

Random forest is a Supervised Machine Learning Algorithm that is used widely in Classification and Regression problems. It builds decision trees on different samples and takes their majority vote for classification and average in case of regression. One of the most important features of the Random Forest Algorithm is that it can handle the data set containing continuous variables as in the case of regression and categorical variables as in the case of classification. It performs better results for classification problems.

- ## LOGISTIC REGRESSION:

Logistic regression is one of the most popular Machine Learning algorithms, which comes under the Supervised Learning technique. It is used for predicting the categorical dependent variable using a given set of independent variables.

Logistic regression predicts the output of a categorical dependent variable. Therefore the outcome must be a categorical or discrete value. It can be either Yes or No, 0 or 1, true or False, etc. but instead of giving the exact value as 0 and 1, it gives the probabilistic values which lie between 0 and 1.

- ## XG BOOST:

XGBoost is an open-source Python library that provides a gradient boosting framework.It helps in producing a highly efficient, flexible, and portable model.

XGBoost is an implementation of gradient boosted decision trees designed for speed and performance.

In prediction problems involving unstructured data (images, text, etc.) artificial neural networks tend to outperform all other algorithms or frameworks.

HIST GRADIENT BOOST CLASSIFIER:

Gradient boosting is an ensemble of decision trees algorithms.Gradient boosting is a generalization of boosting algorithms like AdaBoost to a statistical framework that treats the training process as an additive model and allows arbitrary loss functions to be used, greatly improving the capability of the technique.

As such, gradient boosting ensembles are the go-to technique for most structured (e.g. tabular data) predictive modeling tasks.

RESULTS

DEPLOYMENT

Deployment is the method by which you integrate a machine learning model into an existing production environment to make practical business decisions based on data. It is one of the last stages in the machine learning life cycle and can be one of the most cumbersome.

In order to start using a model for practical decision-making, it needs to be effectively deployed into production. If you cannot reliably get practical insights from your model, then the impact of the model is severely limited.

DEPLOYMENT USING GRADIO:

Gradio is an open-source python library that allows you to quickly create easy-to-use, customizable UI components for your machine learning model.

Gradio allows you to integrate the GUI directly into your Python notebook making it easier to use.

Enables you to create demos of your machine learning model. These demos can be used to present ideas to clients, users, and team members before the actual application is implemented.

CONCLUSION

Through this project we could analyze ,process the data and extract the necessary features to build a model which accurately predicts the winning probability of the first team.

Given the Id’s of teams 1 and 2, applying this model, we can predict the winning probabilities of all teams before the season begins ,using the data of previous seasons.

REFERENCES

- https://www.datacamp.com/community/tutorials/decision-tree-classification-python
- https://www.geeksforgeeks.org/ways-to-import-csv-files-in-google-colab/
- https://gradio.app/docs/