DATA SCIENCE TRAINING CAPSTONE PROJECT REPORT

Name: MANISH R C

Roll No: 202AD127

AIM:

This project's aim is to organize the data and visualize it accordingly. If possible, fit different models for predicting any of another feature.

DATASET: College Football Offensive Stats 2010-2020 | Kaggle

INTRODUCTION:

This dataset revolves around the sport American football which is popularly liked and played during a typical American's college life. The features have different statistical Notion of different teams about their annual scores and parameter that are used to access them. The teams include AirForce, Akron, Alabama, Appalachian State, Arizona, Arizona State, Arkansas, Arkansas State, Army, Auburn, Ball State, Baylor, Wisconsin, Wyoming and many more between the years 2010 and 2020.

Methodology – Overview:

- Gather data: Identify the data sources you will use to answer your business question. Collect the relevant data and store it in a format that is easy to analyze.
- Clean and preprocess data: Once you have gathered your data, you will need to clean and preprocess it. This may involve removing duplicates, filling in missing values, and transforming the data into a format that is easy to analyze.
- Analyze data: Use exploratory data analysis (EDA) techniques to understand your data. This may involve calculating summary statistics, creating visualizations, and identifying any trends or patterns in the data.
- Develop models: Based on the insights you have gained from your data analysis; you can now develop models to answer your business question. This may involve building predictive models, clustering models, or other types of machine learning models.
- Validate models: Validate your models to ensure that they are accurate and reliable. This may involve using techniques such as cross-validation, holdout testing, or A/B testing.
- Communicate results: Once you have validated your models, you can now communicate your results to stakeholders. This may involve creating reports, presentations, or other types of visualizations to help stakeholders understand your findings.

Data Collection and Cleaning:

The dataset is present in Kaggle. Initial impressions are that the records are in the shape of 1372 records with 18 different feature such as,

1. UniversityName-University Name
2. Year-Year of play
3. TeamID-Combination of University name and year
4. CMP-Completed passes in a year.
5. ATT-Attempts to pass in a year.
6. YDS-Amt of yards gained by completed passes.
7. CMPPercent-Percentage of passes successful in a year.
8. YPA-Yards gained on passing divided by total number of passing attempts.
9. LNG-Highest number of yards covered in a single pass of that year.
10. TD-Passing touchdowns in a year
11. INT-Passes that were received by enemy team.
12. SACK-Number of times that team was sacked in that year.
13. SYL-Yards lost from due to sack.
14. RTG-Overall team rating
15. R_ATT-Number of rush play attempts.
16. R_AVG-Avg yards scored during rush play.
17. TOTAL_PLAYS-Total number of passing counts plus rush play counts (offensive play actions).
18. RUN_PERCENT-Percentage of plays that are rush plays.

But from looking at their description using “df.info()”, the following conclusion were made even clear:

- The actual feature names are present in the first row and all the columns are left unnamed except the first one which is also wrong.
- All the above features are present in the object d-type.
- The Team ID feature is not of much use and can be recreated when needed.

Hence the following actions were performed:

- The columns were renamed.
- The first record carrying the feature names were dropped.
- The “TeamId” feature was dropped.
- Each of the column was converted from object datatype to int or float as per their readings present i.e., if contains decimal value then to float and else just a whole number then to integer with necessary needed formatting.

EDA and Visualization:

Firstly, we must identify the outliers present in the data. So, here we plotted a boxplot based on 'RUN_PERCENT'. Then using it, identified the interquartile range. Later we plotted about each teams 'TOTAL_PLAYS' in the year 2020.

I later picked just a particular university (in my case Kentucky University) and analyzed about the performance for the past ten years.

Prediction and Model Building:

I planned on using two specific models Linear Regression and Random Forrest to find how well they fit for predicting the performance of teams. So here I took Kentucky university data and worked with them.

The training dataset had the performance data of the team between 2010 and 2018. I tested for the model to predict for the years 2018 to 2020.

Initially applied linear regression by passing 'CMP', 'YDS', 'YPA','R_ATT', 'R_AVG' of the team as independent variables and 'RUN_PERCENT' as the dependent variable.

Similarly for Random Forrest, all the features were passed except 'RUN_PERCENT' as independent variable and 'RUN_PERCENT' as dependent variable.

Results and Discussion:

I took the mean-squared error as the performance parameter and got 1.8 and 2.25 as the errors for Linear Regression and Random Forrest respectively.

Conclusion:

Hence, an Analysis of the college dataset has been performed and fit into different models and their performances have been identified.

References:

pandas documentation — pandas 1.5.3 documentation (pydata.org)

NumPy Documentation

Matplotlib documentation — Matplotlib 3.7.1 documentation

scikit-learn: machine learning in Python — scikit-learn 1.2.2 documentation