Machine Learning Assignment

PROJECT REPORT

TEAM ID: 18

NFL Play Yardage Regression

Name	SRN
Charan K	PES2UG23CS145
Gauthamdev R Holla	PES2UG23CS197

Problem Statement

NFL Play Yard Regression

Focus on predicting yards gained on NFL rushing plays using metadata, including player lineup, formation, and situational variables like weather. The project involves data cleaning, feature engineering, and applying regression models. Performance evaluation is conducted with RMSE and domain-specific metrics to assess predictive viability. The dataset captures detailed NFL play-by-play statistics, offering insights into player and tactical effectiveness in game scenarios.

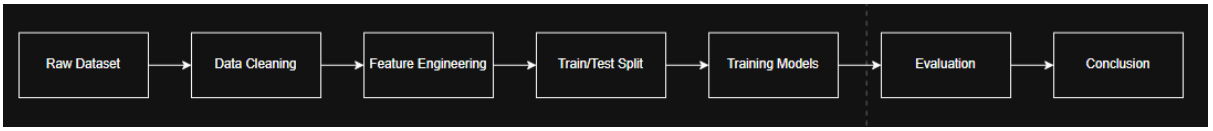
Objective / Aim

- Develop an ML model that accurately predicts yards gained on NFL rushing plays.
- Analyze which factors influence the outcome of a play.
- Evaluate the model's performance using regression metrics like RMSE, MAE, and domain-specific accuracy measures (Ex: Percentage of predictions within ±1/2/4 yards).

Dataset Details

- **Source:** Kaggle – NFL Big Data Bowl 2020
- **Size:** ~682,000 play records and 49 features
- **Key Features:**
    - ***Game Situation:*** Down, Distance, Quarter, YardLine
    - ***Formation and Lineup:*** DefendersInTheBox, PlayDirection, Team
    - ***Weather and Environmental Data:*** Temperature, Humidity, WindSpeed, StadiumType, Turf
- **Target Variable:** Yards – number of yards gained on the rushing play

Architecture Diagram



Methodology

- Imported the NFL Big Data Bowl dataset in Google Colab.
- Data Cleaning was done by removing irrelevant attributes, filling in missing numeric values, and replacing missing categorical values with “Unknown”.
- Created new derived features for detecting patterns.
- Split the dataset into 80% training and 20% testing sets.
- Trained three regression models: Linear Regression (baseline), Random Forest Regressor, and LightGBM Regressor (gradient boosting).
- Used RMSE, MAE, and R^2 and domain metrics for analysis.
- Plotted graphs and analysed features.

Results & Evaluation

Model	RMSE	MAE	R Squared	Within ±1 yd	Within ±2 yd	Within ±4 yd
Linear Regression	6.4	3.765	0.017	20.3%	39.5%	69.7%
Random Forest	0.759	0.061	0.986	98.6%	99.1%	99.6%
LightGBM	5.662	3.518	0.230	21.7%	42%	71.4%

- The Random Forest model achieved the best overall performance, with very low RMSE and high R^2 (0.986).
- Linear Regression struggled due to high variance and non-linear relationships.
- LightGBM performed better than Linear Regression but underperformed compared to Random Forest, likely due to limited hyperparameter tuning.
- Situational factors, such as down, Distance, and Turf Type, significantly influenced play outcomes.
- Domain-specific evaluation metrics (± yard accuracy) provide more meaningful insight than generic regression metrics alone.

Conclusion

This project successfully implemented regression models to predict yards gained in NFL rushing plays using metadata. Through cleaning, feature engineering, and model comparison, the Random Forest model emerged as the most effective.