| **Project Title:Horse Race Prediction** |
| --- |
| **Technologies: Data Cleansing, EDA, Visualization, Machine Learning, Tableau/Power-BI** |
| **Domain: Sports** |



**1. Introduction**

This project aims to predict horse racing outcomes using machine learning techniques. The dataset includes detailed information on horse races and individual horses from 1990 to 2020. Given the complexity and inherent unpredictability of horse racing, this project seeks to explore various machine learning models and feature engineering techniques to improve prediction accuracy.

**2. Dataset Description**

The dataset consists of two main types of files: races and horses, each available for every year from 1990 to 2020. Additionally, a forward.csv file contains information collected before the race starts.

[**Dataset Link**](https://drive.google.com/drive/folders/1RoO0Z_RDBZBQEgakkuDzUB_NiYEV0NJv?usp=sharing), [**Column Description**](https://docs.google.com/spreadsheets/d/1Ql4BqlDUICcjvQba-dZulnCoY3D7tFZYAAw7sXGd6CQ/edit?usp=sharing), [**Forward Data (forward.csv**](https://drive.google.com/file/d/1gkuemJ4v1mlcX4_rH8hKbB10IFEVP_EW/view?usp=sharing)**):** Contains information collected prior to race starts, including average odds from Oddschecker.com, and current RPR and TR values.

**3. Project Goals**

* Primary Goal: To predict the outcome of horse races (e.g., win or place).
* Secondary Goals:
  + To identify significant features affecting race outcomes.
  + To explore the imbalanced nature of the dataset and develop techniques to handle it.
  + To create a robust prediction model using historical data.

**4. Data Preprocessing**

* Data Cleaning:
  + Handle missing values.
  + Normalize data where necessary (e.g., times, distances).
  + Convert categorical variables to numerical representations (e.g., encoding country codes, race conditions).
* Feature Engineering:
  + Create new features based on existing data (e.g., performance metrics from past races).
  + Aggregate features across multiple races to capture trends.
* Data Integration:
  + Merge race and horse datasets on rid to create a comprehensive dataset for analysis.

**5. Exploratory Data Analysis (EDA)**

* Descriptive Statistics:
  + Summary statistics of key features.
  + Distribution plots of continuous variables.
* Correlation Analysis:
  + Correlation matrix to identify relationships between features.
  + Feature importance analysis using mutual information and other techniques.
* Visualization:
  + Scatter plots, histograms, and box plots to visualize data distribution.
  + Heatmaps for correlation visualization.

**6. Modeling Approach**

* Model Selection:
  + Evaluate various machine learning models (e.g.,Regression, Random Forest, Gradient Boosting, Neural Networks).
  + Use cross-validation to assess model performance.
* Handling Imbalanced Data:
  + Use techniques such as SMOTE (Synthetic Minority Over-sampling Technique), under-sampling, and class weight adjustments.
* Feature Selection:
  + Recursive Feature Elimination (RFE).
  + Regularization techniques to reduce model complexity and prevent overfitting.
* Hyperparameter Tuning:
  + Grid search and random search for optimal hyperparameters.

**Note:**

After completion of all the task you need to create a PowerPoint presentation that should contain the:

1. Problem Statement

2. Tools Used

3. Approaches

4. EDA Insights

5. Conclusion/Suggestion

**Project Evaluation metrics:**

* Project evaluation will be done in the live session and have to showcase the approaches done to complete the project
* You are supposed to write a code in a modular fashion (in functional blocks)
* Maintainable: It can be maintained, even as your codebase grows.
* Portable: It works the same in every environment (operating system)
* You have to maintain your code on GitHub.(Mandatory)
* You have to keep your GitHub repo public so that anyone can check yourcode.(Mandatory)
* Proper readme file you have to maintain for any project development(Mandatory)
* Follow the coding standards: <https://www.python.org/dev/peps/pep-0008/>
* You should include basic workflow and execution of the entire project in the readme file on GitHub