Predicting UK Racehorses' Risk of Injury Based on Historical Race and Biometric Data

Sabrina del Rosal

Topic: Predicting UK Racehorses' Risk of Injury and Performance Decline

- **Problem**: Racehorses undergo intense physical strain during races, leading to frequent injuries and performance decline.
- **Opportunity**: By predicting injuries, trainers, jockeys, and owners can optimize training schedules, reduce injury risks, and improve performance.
- **Affected Stakeholders**:
    - **Horse Trainers**: Better training management.
    - **Jockeys**: Understand horse performance on race day.
    - **Owners**: Reduce costs and increase success rates.

Introduction to the Dataset, Data Quality Concerns, and Preliminary EDA

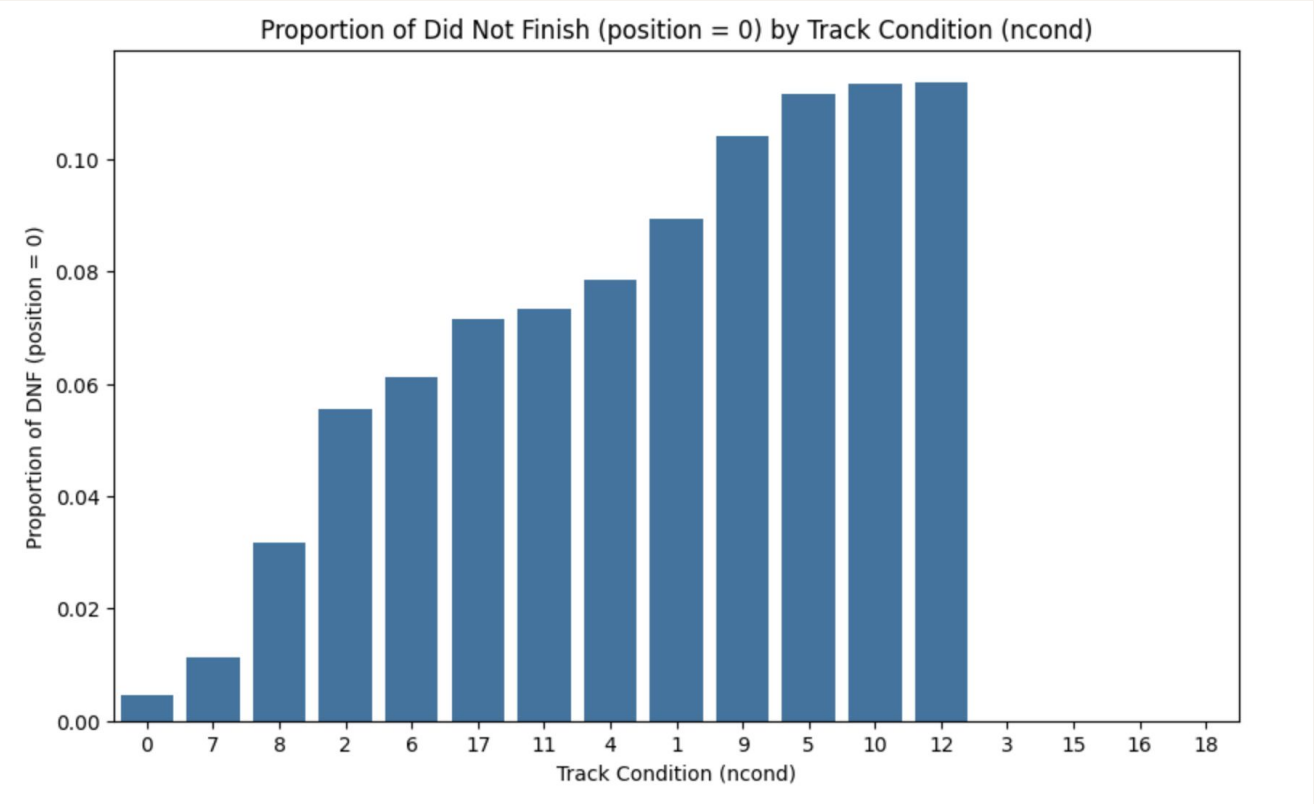
Dataset Overview:

- Pre-Race Parameters: Information collected before the race (odds, trainer data, horse statistics)
- **Race Parameters**: Data on race conditions, times, and prize money.
- **Horse Parameters**: Horse-specific data (weight, jockey, race positions, historical performance)

Data Quality Concerns:

- **Missing Data**
- **Inconsistent Data**: Variability in how data is recorded and labelled
- Both Race & Horse **csv files separated by years** (which years to choose and how many)

Preliminary Findings from EDA:



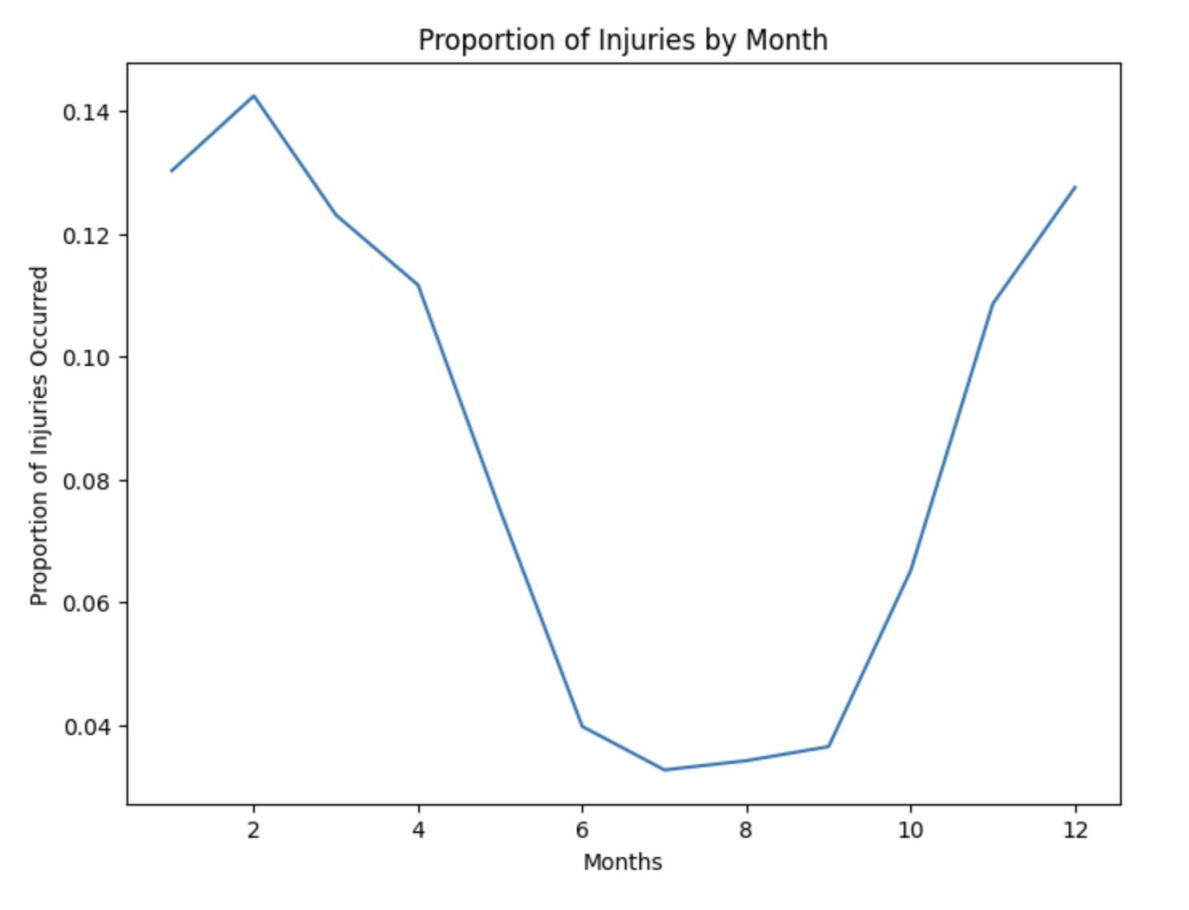
11% DNF for conditions :

12 soft to heavy

10 good to soft

5 soft condition

"Horses with sensitive **tendons**, will feel more at ease on **hard surfaces** because they facilitate their locomotion. Other horses, such as those with **joint** problems, will move more smoothly on **soft surfaces**."



Winter Months:

1. Muscle Stiffness
2. Reduced Flexibility
3. Slippery
4. Longer Warm-Ups
5. Reduced Training

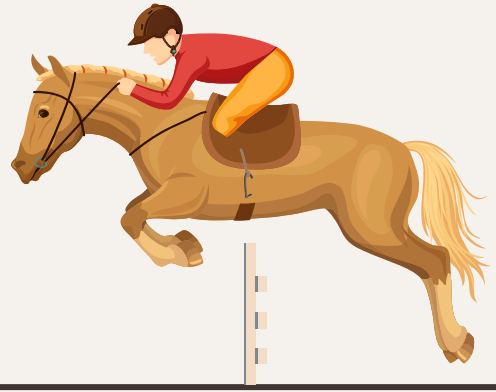
Baseline Models & Evaluations:

- **Objective 1**: Develop a model to classify injury risk by analyzing performance trends and biometric data.

1. Logistic Regression (injury risk classification)

2. Random Forest (injury risk classification)

Model	Training Accuracy	Test Accuracy	Precision (0/1)	Recall (0/1)	F1-Score (0/1)	Macro Avg (F1)	Weighted Avg (F1)
Basic Logistic Regression	0.8262	0.9229	0.93 / 0.57	0.99 / 0.09	0.96 / 0.16	0.56	0.90
LogReg 2 (Grid Search C Params)	0.7191	0.7568	0.98 / 0.23	0.75 / 0.86	0.85 / 0.36	0.60	0.81
Standard Random Forest	0.99998	0.95168	0.96 / 0.76	0.98 / 0.57	0.97 / 0.65	0.81	0.95
Random Forest 2 (Grid Search Params)	0.9234	0.9532	0.96 / 0.78	0.99 / 0.56	0.97 / 0.65	0.81	0.95



Next Steps:

- **Adjust class weights** to give even more importance to injuries.
- Work with other models like **XG Boost** or including **SMOTE** to handle class imbalance.

Main Takeaway:

My goal for this model is to be able to predict and thus **prevent horse injuries**. Because of this, I want a model that **maximizes its' positive rates** even if that means the losing some accuracy by classifying low-risk horses as high-risk.

LogReg 2 (Grid Search C Params = 0.01)

Thank You!