NBA PLAYER INJURY ANALYSIS & FUTURE IMPLICATION

DSO 579 - ADVANCED SPORTS ANALYTICS

TEAM RIVAS II

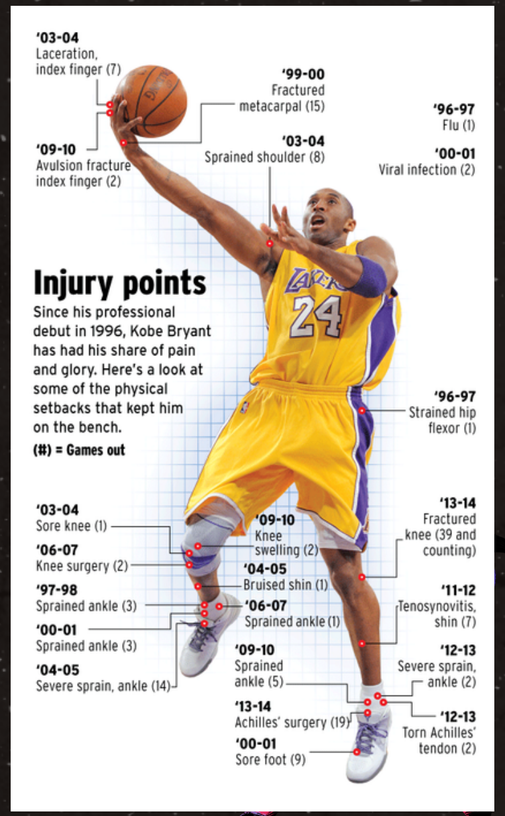
NATHAN CHEOU
JENNY JIN
ALBERTO RIVAS II
JECO CHEUNG
ABED KASSEM

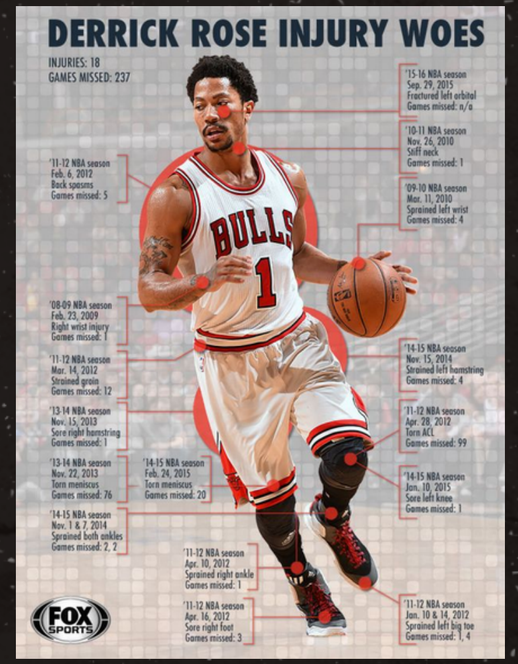
OVERVIEW

01 INTRO
02 DATASETS
03 EDA
04 ML MODELS
05 CONCLUSIONS
06 DISCUSSION

INTRO

THE PURPOSE OF THIS ANALYSIS IS TO EXPLORE INJURIES IN THE NBA. WE WANT TO LOOK AT HOW FREQUENTLY THEY OCCUR, WHAT TYPES ARE THE MOST COMMON, AND MOST IMPORTANTLY, WHAT FACTORS ARE IMPORTANT IN PREDICTING IF AN INJURY WILL OCCUR?





DATASETS

INACTIVE LIST

	Date	Team	Acquired	Relinquished	Notes
0	2010-10-03	Bulls	NaN	Carlos Boozer	fractured bone in right pinky finger (out inde...
1	2010-10-06	Pistons	NaN	Jonas Jerebko	torn right Achilles tendon (out indefinitely)
2	2010-10-06	Pistons	NaN	Terrico White	broken fifth metatarsal in right foot (out ind...
3	2010-10-08	Blazers	NaN	Jeff Ayres	torn ACL in right knee (out indefinitely)
4	2010-10-08	Nets	NaN	Troy Murphy	strained lower back (out indefinitely)

PLAYER STATISTICS

	PLAYER_ID	PLAYER_NAME	SEASON	SEASON_NUM	AGE	PLAYER_HEIGHT_INCHES	PLAYER_WEIGHT	GP	MIN	USG_PCT	...	AVG_SEC_PER_TOUCH	AVG_DRIB_PER_TOUCH	ELBOW_TOUCHES	POST_TOUCHES	PAINT_TOUCHES	TEAM	INJURED ON	RETURNED	DAYS MISSED	INJURED_TYPE
1851	203932	Aaron Gordon	19-20	19.5	24.0	80	235	62	32.5	0.205	...	2.88	1.91	1.8	3.2	4.3	NaN	NaN	NaN	NaN	NaN
1852	1628988	Aaron Holiday	19-20	19.5	23.0	72	185	66	24.5	0.182	...	4.35	4.17	0.1	0.0	0.3	NaN	NaN	NaN	NaN	NaN
1853	1627846	Abdel Nader	19-20	19.5	26.0	77	225	55	15.8	0.164	...	2.05	1.41	0.2	0.0	0.4	Thunder	2020-01-15	2020-01-29	14.0	Sprained_ankle
1854	1629690	Adam Mokoka	19-20	19.5	21.0	77	190	11	10.2	0.110	...	1.47	0.83	0.2	0.0	0.8	NaN	NaN	NaN	NaN	NaN
1855	1629678	Admiral Schofield	19-20	19.5	23.0	77	241	33	11.2	0.118	...	1.40	0.51	0.2	0.2	0.6	NaN	NaN	NaN	NaN	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
5573	2584	Willie Green	13-14	13.5	32.0	75	201	55	15.8	0.162	...	2.44	1.84	0.5	0.1	0.5	NaN	NaN	NaN	NaN	NaN
5574	201163	Wilson Chandler	13-14	13.5	27.0	80	225	62	31.1	0.193	...	2.33	1.26	1.4	1.0	1.7	NaN	NaN	NaN	NaN	NaN
5575	202333	Xavier Henry	13-14	13.5	23.0	78	220	43	21.1	0.223	...	3.24	2.45	0.9	0.1	1.2	NaN	NaN	NaN	NaN	NaN
5576	2216	Zach Randolph	13-14	13.5	32.0	81	260	79	34.2	0.258	...	2.15	0.81	3.6	17.3	9.5	NaN	NaN	NaN	NaN	NaN
5577	2585	Zaza Pachulia	13-14	13.5	30.0	83	275	53	25.0	0.167	...	1.83	0.54	9.1	3.0	6.3	NaN	NaN	NaN	NaN	NaN

DATASETS

1. INJURY CLASSIFICATION

2. SEASON ENDING INJURIES

3. REMOVING DUPLICATES

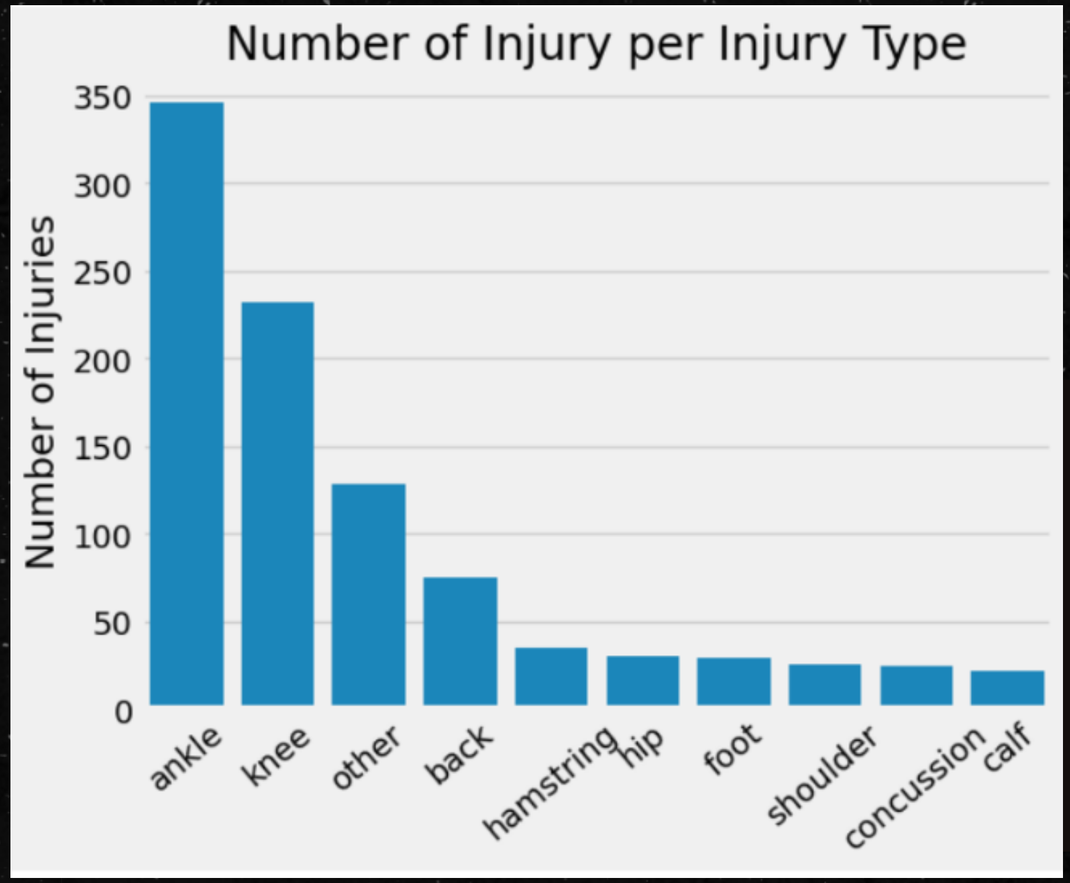
4. COMBINING DATASETS

EXPLORATORY ANALYSIS

1. INJURIES BY TYPE
2. LOWER BODY INJURIES BY WEIGHT AND HEIGHT
3. INJURIES BY MONTH AND YEAR
4. INJURIES TAKING PLAYERS OUT FOR THE SEASON

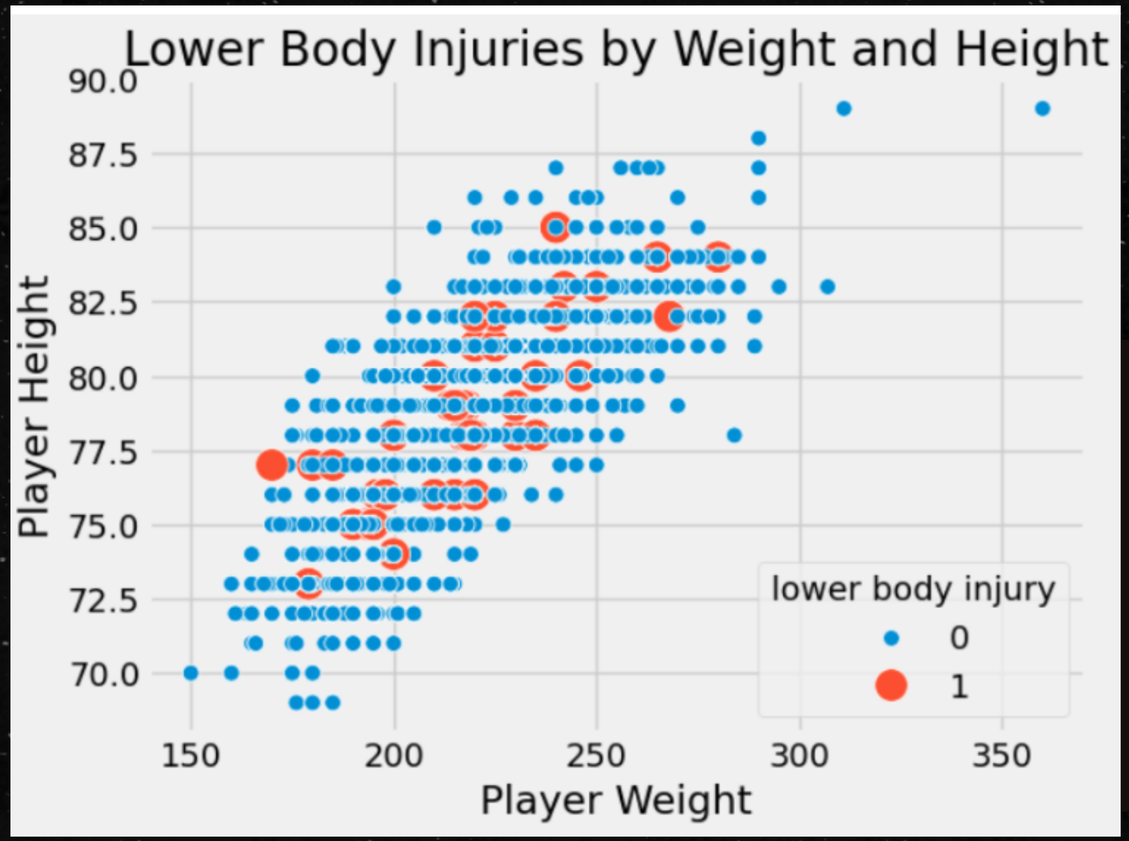
INJURIES BY TYPE

MOST COMMON INJURIES WERE ANKLE AND KNEE

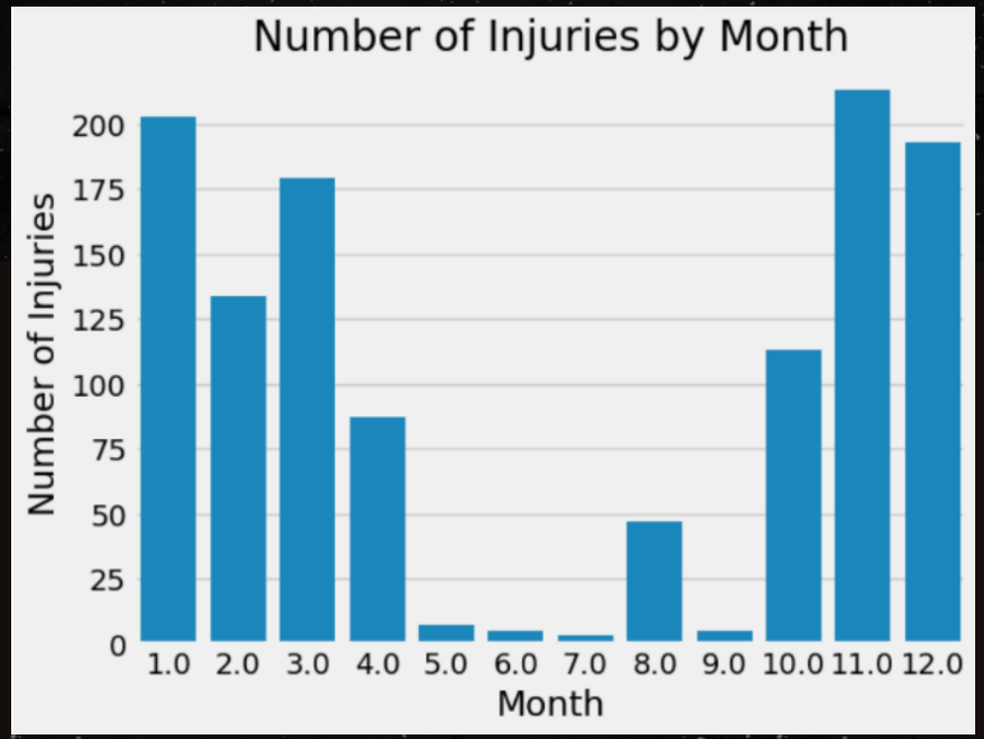


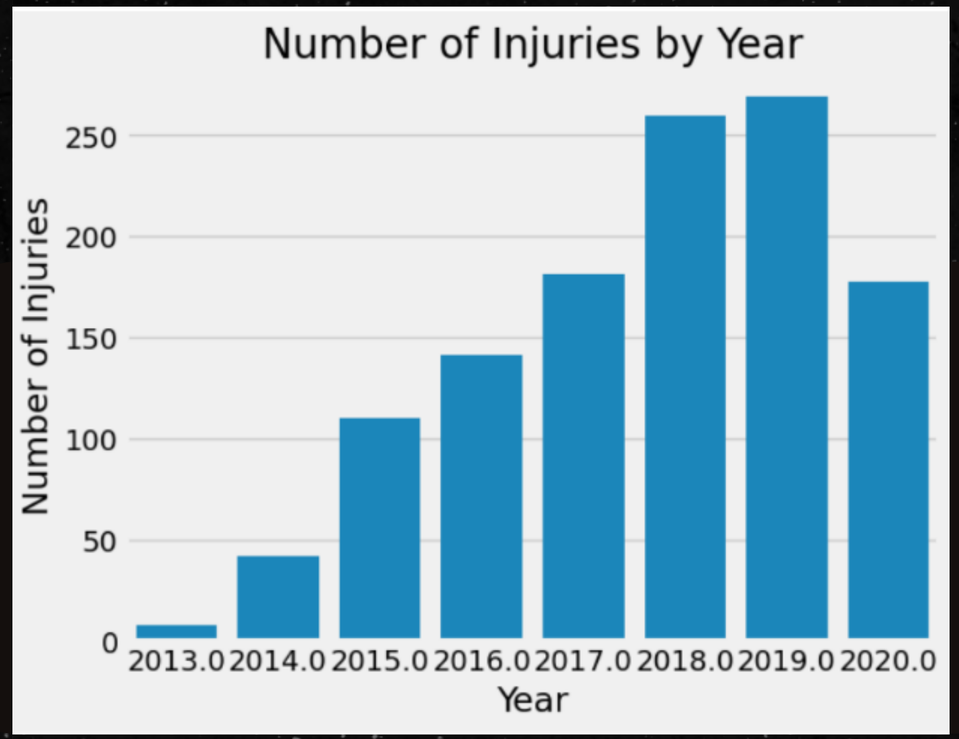
LOWER BODY INJURIES BY WEIGHT AND HEIGHT

LOWER BODY INJURIES REFER TO ANKLE AND KNEE INJURIES, WHERE MUCH OF THE PLAYER'S WEIGHT IS EXERTED



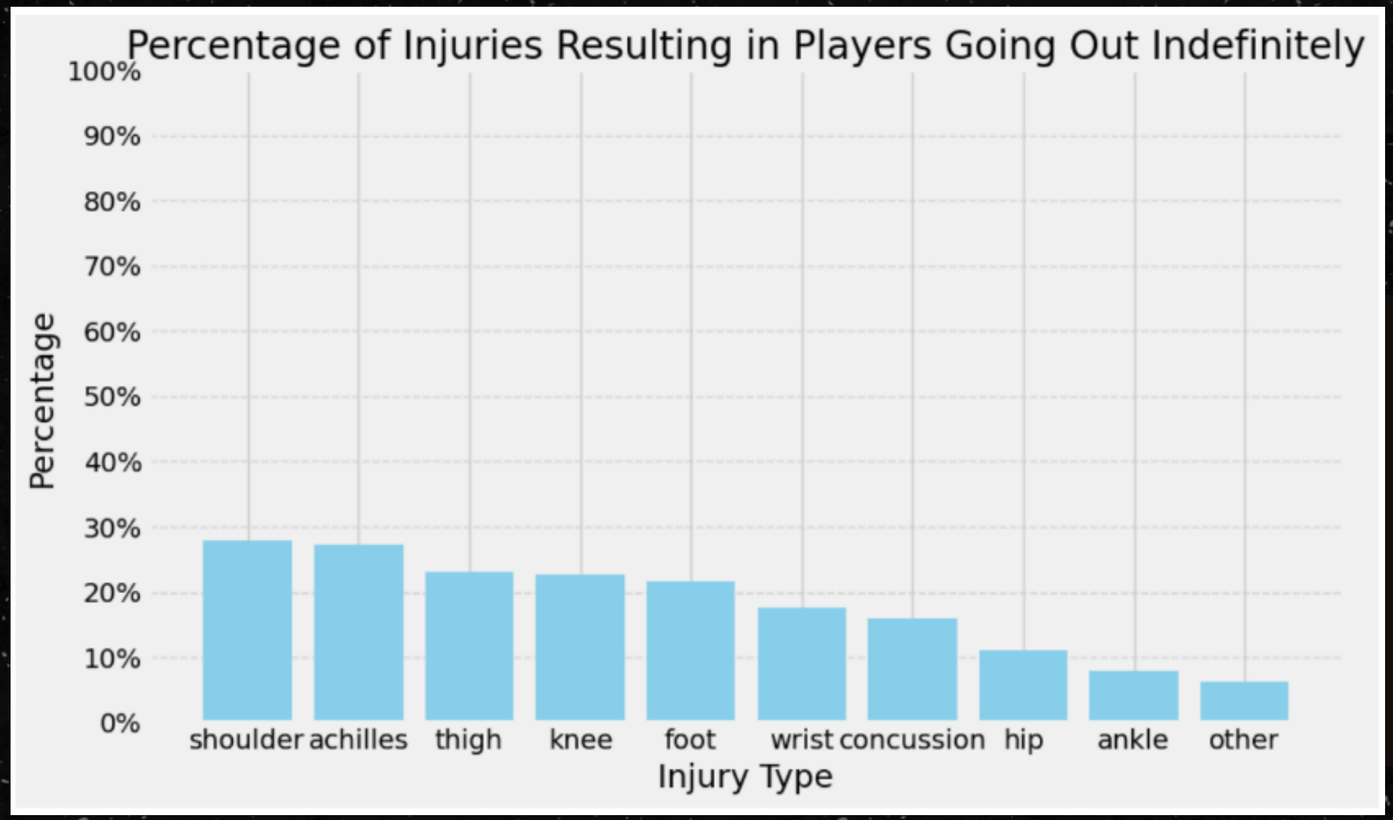
INJURIES BY MONTH AND YEAR





INJURIES TAKING PLAYERS OUT FOR THE SEASON

SHOULDER INJURIES, THE 8TH MOST COMMON INJURY



MODEL BUILDING

1. Feature Selection
2. Logistic Regression
3. Decision Tree Classification with Ensembles

FEATURE SELECTION

In our final dataset, we have 4274 observations and 37 columns. It is problematic to include everything while building our model, due to the low interpretability (overly complex model) and high computational complexity. Therefore, feature selection is crutial in this case.

STEPWISE METHOD

- Definition: A popular data-mining tool that uses statistical significance to select the explanatory variables to be used in a multiple-regression model.

FINAL

- By running the stepwise regression, we ended up with 9 features.
- Resulting features: ['DIST_MILES', 'PACE', 'PAINT_TOUCHES', 'AGE', 'POSS', 'AVG_SPEED', 'GP', 'DRIVE_FGA', 'PLAYER_HEIGHT_INCHES']

LOGISTIC REGRESSION

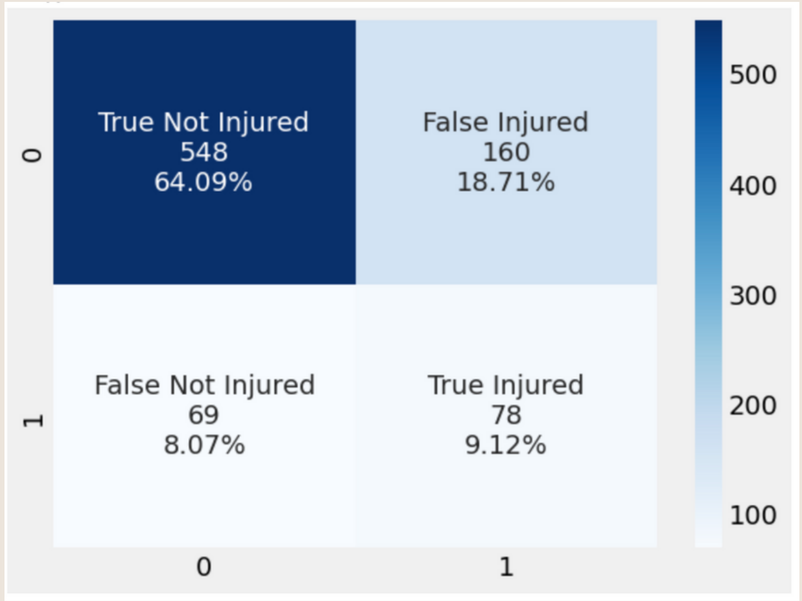
```
Optimization terminated successfully.
         Current function value: 0.489443
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                Injured   No. Observations:                 3419
Model:                          Logit   Df Residuals:                     3409
Method:                           MLE   Df Model:                            9
Date:                Sun, 25 Feb 2024   Pseudo R-squ.:                  0.1721
Time:                        03:47:00   Log-Likelihood:                -1673.4
converged:                       True   LL-Null:                       -2021.3
Covariance Type:            nonrobust   LLR p-value:                5.514e-144
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const        -15.5382      2.201     -7.060      0.000     -19.852     -11.225
DIST_MILES     2.3752      0.214     11.125      0.000       1.957       2.794
PACE           0.1006      0.012      8.707      0.000       0.078       0.123
PAINT_TOUCHES  0.0861      0.026      3.367      0.001       0.036       0.136
AGE           -0.0463      0.011     -4.028      0.000      -0.069      -0.024
POSS          -0.0006      0.000     -5.247      0.000      -0.001      -0.000
AVG_SPEED     -0.6781      0.204     -3.325      0.001      -1.078      -0.278
GP             0.0251      0.006      4.194      0.000       0.013       0.037
DRIVE_FGA      0.1240      0.037      3.334      0.001       0.051       0.197
PLAYER_HEIGHT_INCHES  0.0600      0.019      3.218      0.001       0.023       0.097
==============================================================================
```

STATSMODEL RESULT

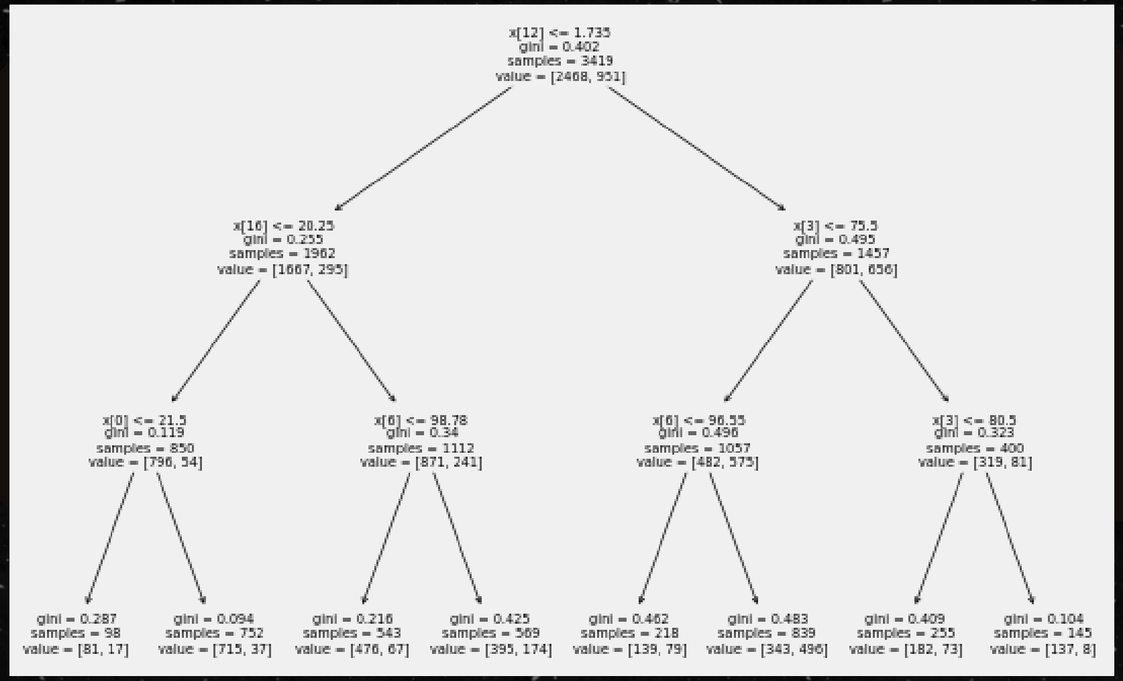
- Attached is our Logistic Regression statsmodels result using the 9 variables selected in the stepwise regression.
- All the variables have p-values under 0.05

CONFUSION MATRIX

- **Logistic Regression Score: 73.22%**



DECISION TREE CLASSIFIER



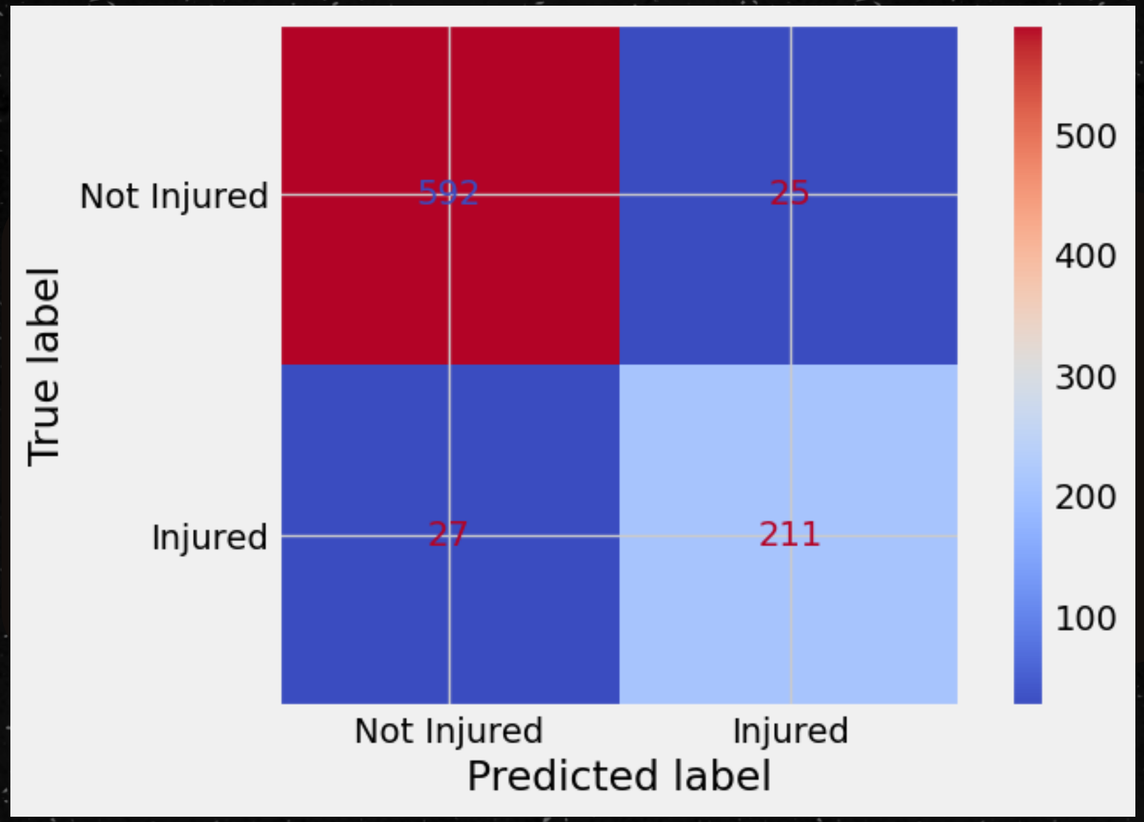
GRIDSEARCHCV

- Used GridSearchCV to find optimal hyperparameters
- criterion (Gini), max_depth (3), min_samples_leaf (4), min_samples_split (2)

RESULTS

- Accuracy Score: ~75.17%
- X[12] = distance in miles
- x[3] = games played

ENSEMBLE METHODS



METHODS USED

- Bagging, Random Forest, XGBoost Clasifiers
- Wanted to compare ensemble methods to basic decision trees and classifiers

RESULTS

- Bagging Ensemble: 93.92%
- XGBoost: 93.33%
- Base Decision Tree:75.17%
- Random Forest: 75.67%
- Logistic Regression: 73.22%

IMPLIMENTATION

Starting Point For Injury Prevention Strategies

FUTURE IMPROVEMENT

COLLABORATION WITH EXPERTS:

trainers, coaches, and biomechanics experts for domain expertise to tailor prevention protocols

INTEGRATION OF PHYSIOLOGICAL DATA:

such as muscle strength, flexibility, and biomechanical factors from wearable sensors

EXPANDED INJURY CLASSIFICATION:

include a wider range of injury types and severity levels

NATURAL LANGUAGE PROCESSING:

extract and categorize injury information from unstructured text data more effectively

LIMITATION

Limited Data Duration:

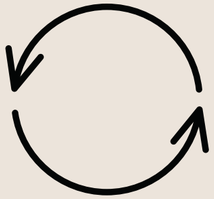
Only 7 years of data available

Duplicate Injury Methodology:

No account for multiple injuries to different body parts in one incident

Lack of Contextual Information:

injury mechanisms, player conditions, or external factors like game intensity



Accuracy and Consistency of Reporting:

Variability in injury reporting across teams and seasons

CONCLUSION

PREVALENT INJURY

- Lower Body and Shoulder
- Model: Distance Covered

RISK AND IMPROVE

- Limitations
- Continuous Imporvement

THANK YOU