MIT SLOAN
SPORTS ANALYTICS CONFERENCE
MARCH 11 - 12, 2016 BOSTON CONVENTION AND EXHIBITION CENTER

Preventing in-game injuries for NBA players

Basketball
Paper ID: 1590

Hisham Talukder**[A], Thomas Vincent**[A], Geoff Foster[B], Camden Hu[B], Juan Huerta[A], Aparna Kumar[A], Mark Malazarte[A], Diego Saldana[A], Shawn Simpson[A]

hisham.talukder@dowjones.com
tlfvincent@gmail.com
** Co-first authors
[A] Dow Jones [B] Wall Street Journal

Abstract

Player injuries have long been a cause of concern to NBA team management and fans, as they can significantly affect the overall performance of a team. Over the past 2 years many high caliber players, such as Kobe Bryant, Kevin Durant and Derrick Rose, have missed significant amounts of playing time as a result of injuries. Consequently, some have argued for drastic changes such as reducing the NBA season or imposing longer layoff periods during the regular season. In this work we present a model that offers a quantitative and systematic approach to injury prevention by allowing teams to predict the likelihood that any given player will succumb to an injury event during the course of an upcoming game.

We apply advanced machine learning techniques to predict the probability of injury for a player. Our model is based on play-by-play game data, SportsVU data, player workload and measurements, and team schedules from the last 2 years. Our results demonstrate strong accuracy in predicting whether a player will get injured in an upcoming week. By combining these results with information on team schedules and rest days, our approach enables team management and decision-makers to identify the best time for a team to rest their star players and reduce the risk of long-term injuries, while optimizing team strategies.

Finally, we show the effect of injuries on NBA teams as well as on the fans' experience. By accounting for the amount of money invested in each player, we can rank player injuries based on the financial cost of missed games associated with these injuries. We anticipate our predictive model to be useful to not only NBA managers and coaches to help understand the optimal time for resting players, but also to the millions of fans that currently participate in the growing fantasy-basketball world. This can be a valuable model for fantasy sports players with regards to preseason rankings and in-season pick ups.

1. Introduction

Over the course of the NBA's history, the outcomes of games and championships have been determined through a mixture of talent, systematic strategy, and luck. A prominent example of an aspect of the sport where things are left to luck is the incidence of injuries to star players, or in more extreme cases, the so-called "injury bug" that occasionally hits teams [6]. Aside from the obvious impact of injuries on rotation and team dynamics, they can also impact revenue generated by advertising, fan attendance and merchandising.

To quantify the impact on revenue, we undertook a historical analysis on the total amount of seasonal team salary losses that can be attributed to player injuries from 2000 through 2015 (Fig 1B). The analysis reveals that seasonal losses during this time frame range from 10-50 million

dollars per NBA team. Despite the substantial negative impact of in-game player injuries, currently, there exists no methodology to systematically predict whether a player is likely to succumb to injury in any upcoming games.

In this work we quantify the net impact of injuries and propose a statistical model to determine when a player is considered at-risk of an injury. Our method has strong performance in terms of predictive accuracy and lift, and enables a team to keep track of the incremental risk of injury over time. We also provide a breakdown the main factors (schedule, burden on the player, playing style, etc.) that contribute to in-game injuries, and provide a clear set of recommendations for when a player should be rested or when playing time should be reduced. By combining these insights with information on team schedules and rest days, our approach provides team management and decision-makers with the ability to identify the optimal times to rest their star players and reduce the risk of long-term injuries.

2. Data Aggregation & Modeling of in-game injuries

2.1: Data collection and aggregation

In order to gather the data needed for our analysis, we applied custom scrapers, open-source libraries and freely available resources to collect, parse and aggregate this data. Injury data was collected from the Pro Sports Transactions website [9]. We removed games missed by players due to the flu, illness, sports hernia, family emergency or any other non- in-game injury related causes. We focused on NBA players who played at least 15 minutes a game on a nightly basis. After cleaning and removing rare injuries, we were left with 500 in-game injuries over the last two years (2013-2014 and 2014-2015 seasons). This data contains start/return dates for all known injury events. In addition to injury event data, we also used game-by-game player statistics for the last two years, which also included SportsVu advanced metrics data over the same time period [11]. Finally, physical player and other characteristics such as height, weight, age and salaries were obtained from Basketball Reference and ESPN [9,10].

2.2: Sliding window technique

We took a sliding window approach [4] (summarized in Figure 2A) to create a data set for modeling. For each day with an NBA game, we create both an associated aggregation time window and a prediction time window. We aggregate all in-game statistics from a fixed time period in the aggregation window. We combine this information with data on player characteristics to create a game-player-level feature matrix for each game. Injury events are our response variable. We measured the response within the prediction window -- i.e. the game is associated with an injury if a player has gotten injured during the 7 days following the game. We ran our model over a wide grid of possible prediction window and aggregation window lengths to determine optimal settings.

2.3: Random Forests to predict injuries

Our method is based on a *committee machine* model [2] that aggregates all in-game statistics over a rolling window of pre-specified length. Empirical evaluation revealed that a 14-day aggregation window and 7 day prediction window provided optimal performance. A random forest with B=100 regression trees was used as a classifier to predict the injury/non-injury status after the 14-day rolling window. A random forest first creates B bootstrap samples from training data, then fits separate classification trees for each bootstrap sample (finding optimal variable/split-points), to finally make a *committee* or *majority vote* on the injury/non-injury status and computing estimated injury probabilities i for each player i based on the average results across the different trees [2,5].

Our training and testing data consisted of a *50%-50%* random splitting (X,Y) from our original input variables:

$$X = (Xminutes, Xdreb, Xfta, Xpts, Xfga, Xblk, Xto, Xstl, Xoreb, Xfg3a, Xspd, Xgame\#, Xb2b, Xassist),$$

with y being the response injury/non-injury indicator after the rolling window. We used multiple statistical validation techniques (described in Section 3) to test the accuracy of our model on the resultant test data after random splitting.

2.4: Variable importance

Alongside the assessment of predictive performance, we explored the relative importance of the P variables included in our model in contributing to the risk of injuries in NBA players. To do this we implemented a procedure in which 90% of the training data was randomly sampled and fitted to the same random forest modeling scheme described in Section 2.3. This was repeated over 20 independent iterations to obtain an overall mean and confidence interval for the relative importance of each variable. This was estimated as follows:

$$imp_k = \frac{\sum_{k=1}^{K} imp\ (j,k)}{K}$$

where $k \in \{1,...,K\}^K$ is the total number of sampling iterations performed, and *imp(j,k)* represents the overall estimated importance of variable $j \in \{1,...,J\}$ as evaluated in iteration k.

3. Results

3.1: Optimizing the aggregation and prediction window sizes

We created a matrix of features using the sliding window framework described in Section 2.2. We applied a grid search to discover the optimal parameter values for the aggregation and prediction window length. Figure 2B shows the average AUC obtained after 10-fold cross-validation on different combinations of prediction and aggregation window sizes [2]. Not surprisingly, we can observe that a longer prediction window allows us to capture injury events more accurately. The usability of the methodology was a key objective of this analysis, so we opted for a prediction and aggregation window size that provided an adequate balance between predictive performance (longer windows) and actionability (shorter windows). In the remainder of this analysis we use a prediction window of 7 days and an aggregation window of 14 days. This combination gives an average AUC of 0.85 based on 10-fold cross validation.

3.2: Predicting the risk of injury

In addition to the construction of our final random-forest based model, we applied an auto-regressive approach with a logistic regression model as a baseline for performance, which yielded an average AUC score of 0.61 based on 10-fold cross-validation. The random forest model as described in Section 3.1 yielded a significant increase in predictive performance with an average AUC of 0.92 (Figure 3A). It is important to note that logistic regression could not achieve similar performance, which seems to suggest that the problem of predicting injuries is not linearly separable. Support-vector machine classifiers yielded results comparable to random forests, but the additional computational complexity of SVMs, along with the lack of interpretability and the necessity of choosing the correct kernel made random forest a more suitable model.

We report ROC curves and associated AUC scores for the 10-fold cross-validation resulting from a random forest model (Figure 3a). We also report the average probabilities, as determined by our model, of injured players versus non injured players in a held out set (Figure 3b). The average *predicted* probability of all players that were not injured in a held out set for a given game was 0.06. This is close to the empirical probability of injury over the last two years. In the same held out set, the average *predicted* probability of injury prior to the game on injury for all players who were actually injured was 0.19. This is a 300% increase in the probability compared to all players. This shows some of the injuries can be prevented by using our model to strategically rest players.

The results obtained above indicate that our model can accurately predict whether a player will suffer an injury in the next 7 days. To further support the validity of this model, we show the lift curve (or cumulative gains chart) obtained from a held out set (Figure 4). The lift curve compares different deciles of the sorted predicted probability (from highest risk of injury to lowest) to the true injured players captured in each decile. For example, if we rested the top 20% of high risk players it would potentially prevent 60% of all injuries (marked in Figure 3). This is a 300% increase compared to randomly selecting 20% of all players to rest. This provides a data driven way to rest players to potentially prevent injuries.

3.3 Identifying individual features that contribute to injury events

We aimed to find the most relevant features that contribute to the risk of injury in NBA players. To do this, we applied the approach described in Section 2.4, which revealed that the five most important features were, by decreasing order, (1) the average speed at which a player ran during games; (2) the total number of games played; (3) the average distance covered by a player; (4) the average number of minutes played; and (5) the average number of field goals attempted. These results are consistent with the expectation that increased individual workload, as demonstrated by the total number of games played, the average number of minutes played and the average number of field goals attempted, is associated with an increased risk of injury. Similarly, playing style, as demonstrated by the average speed at which a player ran during games and the average distance covered by a player, is also associated with increases risk of injury. These findings support the hypothesis that individual workload and playing style can place additional strain on a player's body. Finally, it is interesting to note variables that do not significantly contribute to the prediction of injuries, such as the number of back-to-back games and the number of games played during the 14-day aggregation window used in our model. All together these results indicate that total workload and activity, rather than the short-term increase of games played, are much more substantial factors to the overall risk of injury.

4. Discussion

We have created an accurate model to predict injuries in the NBA. The model provides a game by game viewpoint of the health status of NBA players (who play at least 15 minutes) and their risk injury potential. The usability and accuracy of our model has been demonstrated based on multiple statistical validation techniques. The random forest modeling scheme that was applied in this work was able to reliably detect players at risk of injury, as exemplified by the overall AUC score of 0.86 achieved by our model. The lift curve also validates the usability of our model. It demonstrates that by resting the top 20% of high risk scores at any given day there is a potential to prevent 60% of all injuries. In the NBA currently, players are rested at random times completely decided by the head coach. With our model, there is now a data driven way to strategically rest players throughout the year. Ideally any player that has a probability of getting injured from our model greater than 0.15 should be rested for the next game. With a one game rest the probability of

getting injured decreases drastically. The head coach or general manager can combine the probability with the importance of the opponent to make a more data driven decision.

The NBA recently has changed their schedule to include fewer back to back games to combat the high number of injuries [7]. However, our model shows the number of back to back games in the 14 day aggregation period is one of the least important features in predicting injuries. Instead, the number of total games a player plays does have an effect on injury risk. This means the deeper a player goes into the season the higher the risk of injury. This can also be used to strategically rest players because most of the games missed due to resting should happen towards the second half of the season. Finally, the model shows the style of play is a huge indicator on injuries. The average speed someone runs, the average minutes played and average distance travelled by a player during all games are big indicators of injuries. This is a key factor that should be taken into account to prevent injuries. A coach can start to manage minutes played by the NBA player from the start of the season in a more effective way. Instead of missing entire games due to resting, this can be done over a month by decreasing the number of minutes played each game. Finally, by providing a tool that allows for the strategic rest of NBA players, we expect to provide added financial gains to organizations across the NBA. Indeed, the direct consequence of reducing the number of player injuries incurred over the course of a season would mitigate the financial losses incurred by organizations across the NBA.

5. Conclusions and future work

In this body of work, we have built a predictive model that can be used to strategically rest players at risk of injury. By providing a data-centric process to manage minutes played by NBA players throughout the season, we anticipate our predictive tool to be of great use to NBA managers and coaches alike. While the current work achieved reliable performance, there exists a number of additional features and analysis that could further complement this study. First, the breadth of information provided by SportsVu data could further push our understanding of when and how injuries may occur. For example, statistics on drives to the basket or number of post-ups could unveil further information on susceptibility to injury. Second, we hope to implement this model using a real time framework in order to scale and generate probabilities on a nightly basis. Finally, we plan to expand this analysis to other sports, including MLB, NFL, and soccer.

References

[1] McCullagh, P., Nelder, J.: Generalized Linear Models. Chapman and Hall 1983., London, England (1989).
[2] Hastie, T., Tibshirani, R., and Hastie, T.: The Elements of Statistical Learning. Springer New York (2009).
[3] SJ Wright and J Nocedal. Numerical optimization, volume 2. Springer New York, (1999).
[4] Wang, J. Zivot, E.: Modelling financial time series with S-Plus. Springer New York, (2006)
[5] Efron, B. and Tibshirani, R. J.: An Introduction to the Bootstrap. Chapman & Hall/CRC, London (2006).
[6] Powell, S: Injuries had major impact on playoff outcome this season. http://www.nba.com/2015/news/features/shaun_powell/06/17/5-things-we-learned-from-the-finals/ (2015, June 17)
[7] Zillgitt, J: NBA schedule more player friendly with fewer back to back games. http://www.usatoday.com/story/sports/2015/08/12/nba-schedule-more-player-friendly-fewer-back-back-games/31564149/ (2015, August 13)
[8] Pro Sports Transaction Archive. Pro Sports Transaction Archive. http://www.prosportstransactions.com
[9] Basketball reference. http://www.basketball-reference.com/
[10] ESPN. http://espn.go.com/nba/salaries/_/year
[11] NBA. http://nba.com

Figures

Figure 1A: Total games missed due to player injuries. Distribution of total games missed due to player injuries. Total salary loss was computed on a per-team basis over the entire time period of 2010-2015.

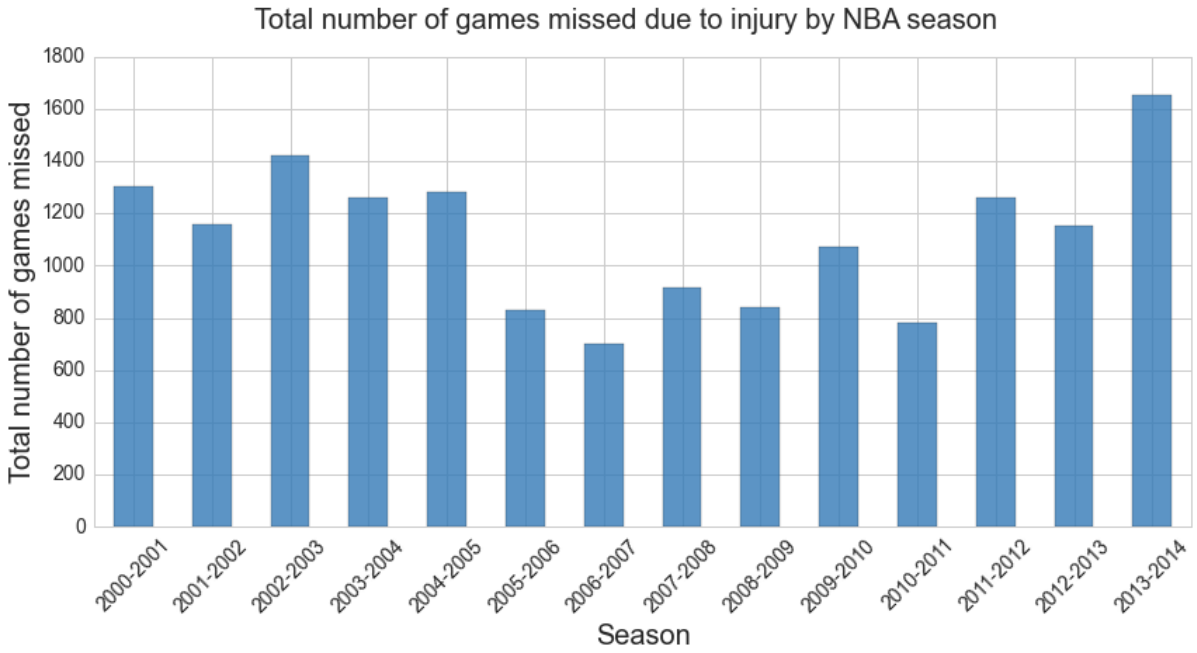

Figure 1B: Total salary loss attributable to player injuries (2000-2015). Distribution of total salary loss attributable to player injuries. Total salary loss was computed on a per-team basis over the entire time period of 2010-2015

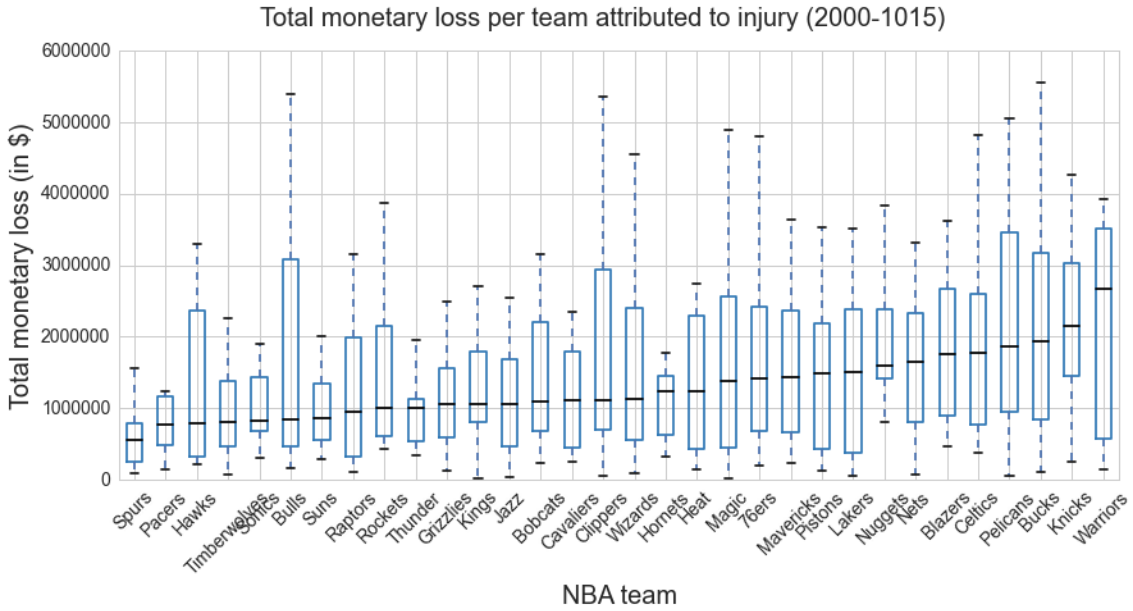

Figure 2A: Illustration of the sliding window approach. This figure displays the sliding window approach used to predict injuries. In this example, all the game in a 14 day span before game day is used to aggregate in game statistics. Combined with physical attribute of NBA players (height, weight, etc.), the X matrix is created. The response variable is whether a player got injured during the 7 day prediction window. This combination of 14 days aggregation window and 7 days prediction window is used for each day where there was an NBA game.

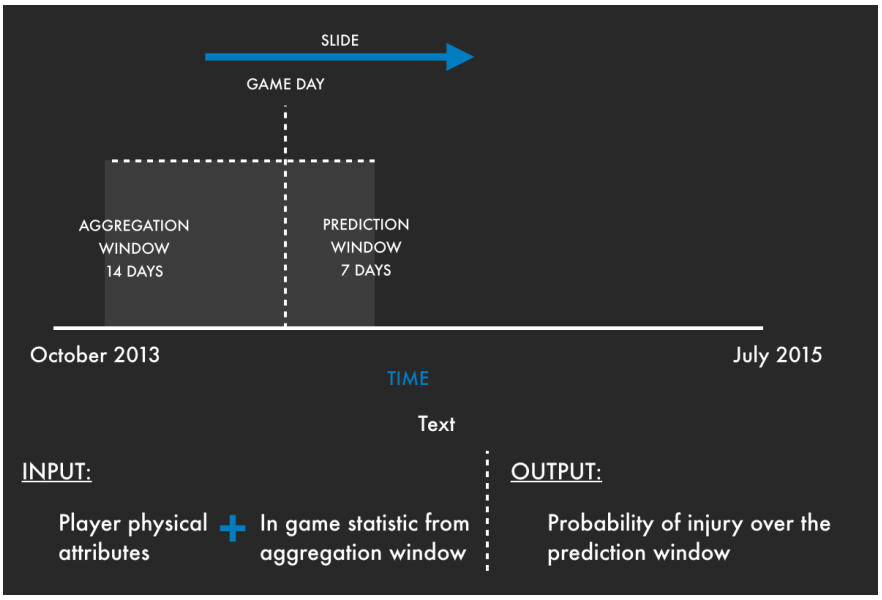

Figure 2B: Average AUC on different window lengths for aggregation and prediction. This plot shows the average AUC on 10 fold cross validation for different levels of aggregation and prediction window. We selected 14 day aggregation window and 7 day prediction window because it resulted in a very good AUC (~.85) while still maintaining a tight prediction window.

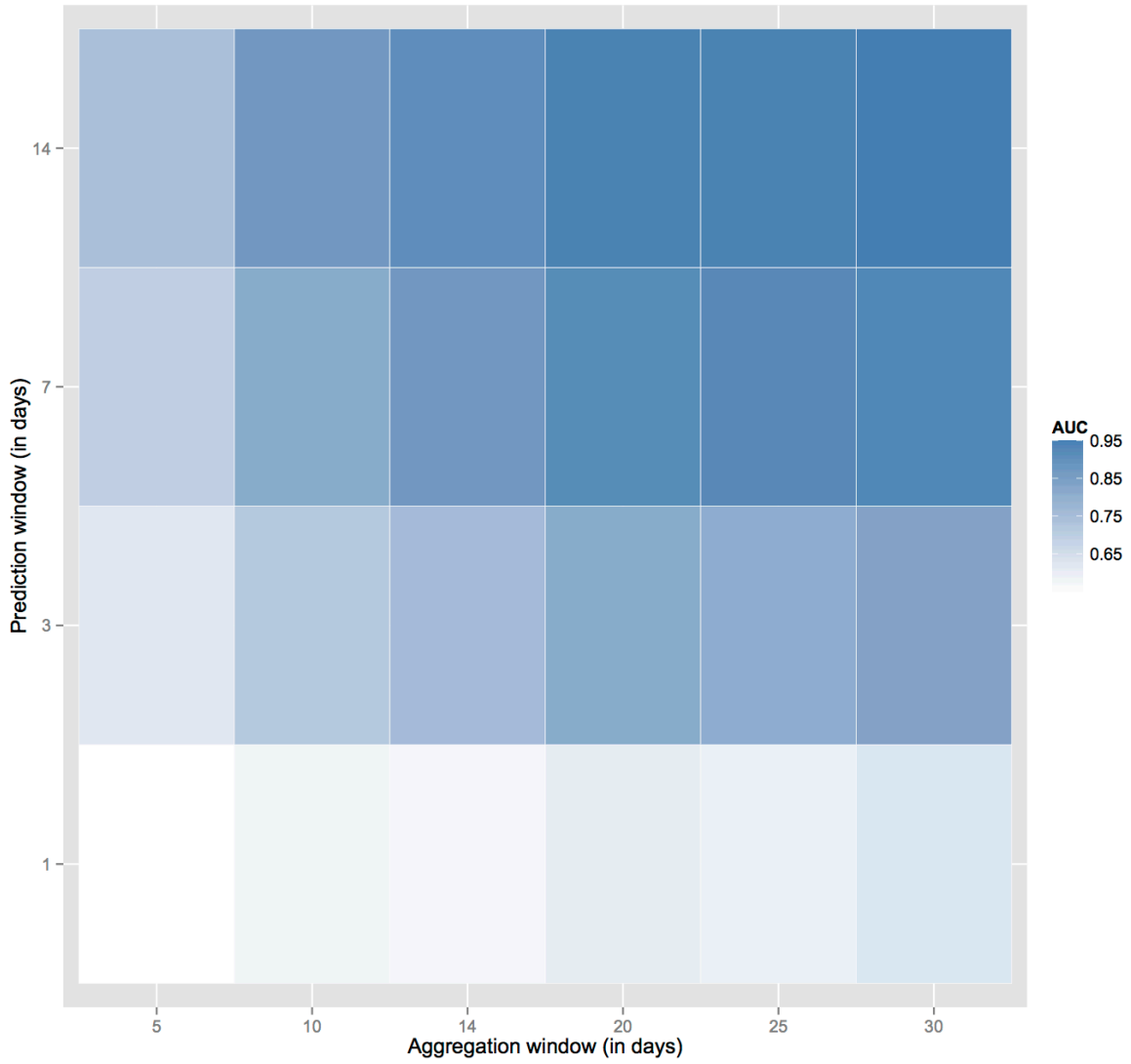

Figure 3a: ROC curve for injury prediction model

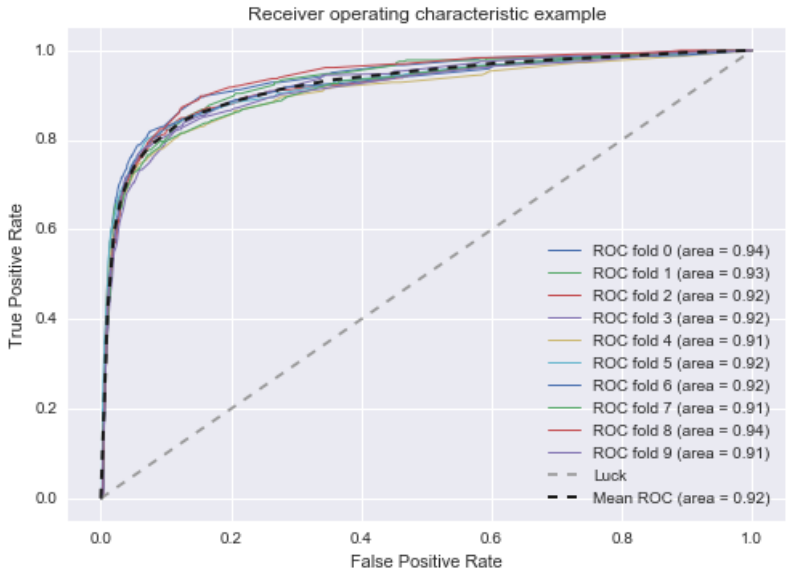

Figure 3b: Distribution of estimated probabilities for 7-day windows that contain an injury event, and for 7-day windows that do not contain an injury event.

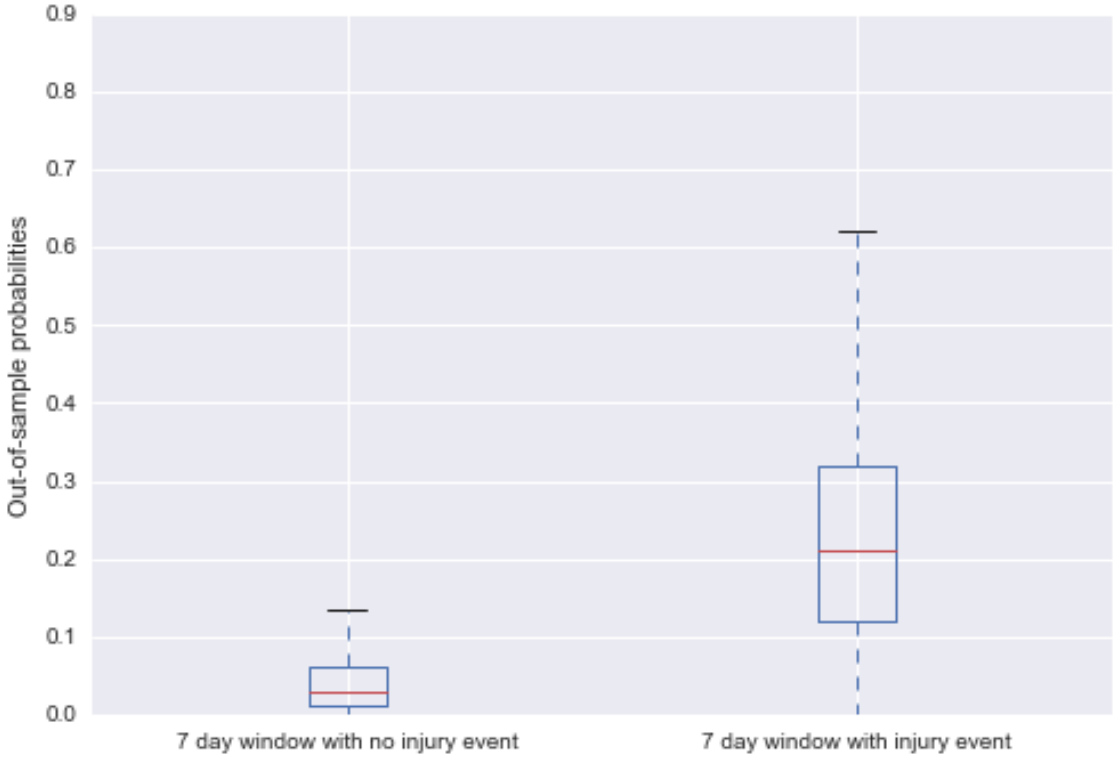

Figure 4: Lift curve for injury prediction model. The lift curve is a way to analyze our model at each percentile of top scores. If we rest the top 20% of injury scores from our model, we can potentially prevent 60% of all injuries. This shows a 300% gain over chance (diagonal line).

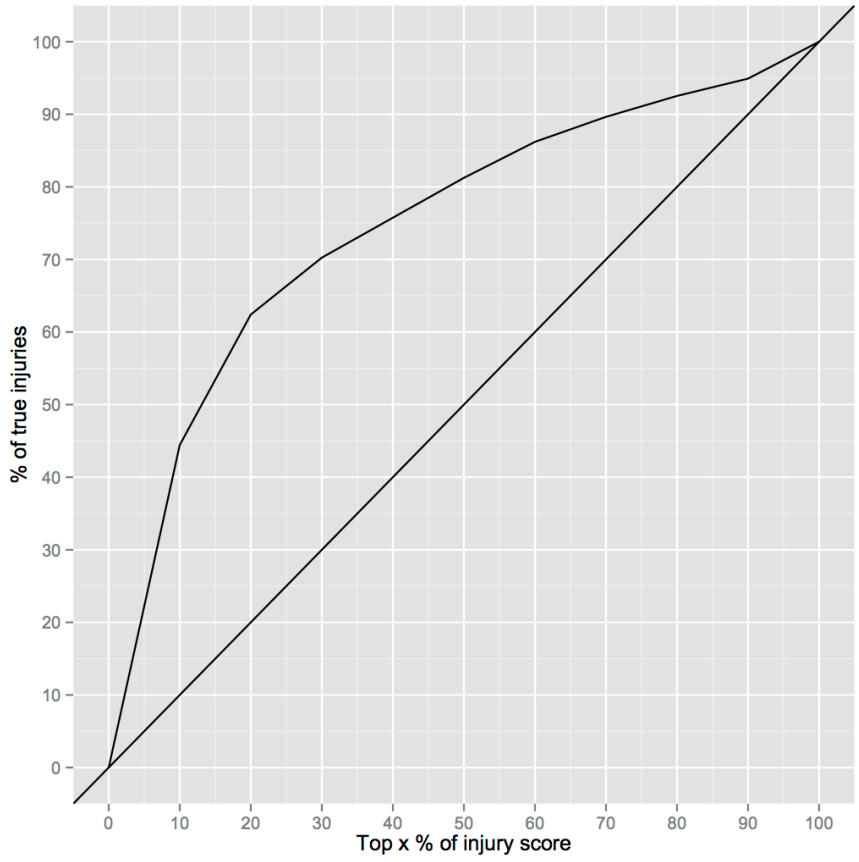

Figure 5: Variable Importance. Here is a ranked list of all variables normalized. The x-axis shows the normalized importance of each variable. For example, average game speed