Injury Prediction in Elite Ice Hockey using Machine Learning

JAKOB CLAESSON, EMIL HÄGLUND, PONTUS STABERG
Injury Prediction in Elite Ice Hockey using Machine Learning

Claesson J, Häglund E, Staberg P

Abstract—Sport clubs are always searching for innovative ways to improve performance and obtain a competitive edge. Sports analytics today is focused primarily on evaluating metrics thought to be directly tied to performance. Injuries indirectly decrease performance and cost substantially in terms of wasted salaries. Existing sports injury research mainly focuses on correlating one specific feature at a time to the risk of injury. This paper provides a multidimensional approach to non-contact injury prediction in Swedish professional ice hockey by applying machine learning on historical data. Several features are correlated simultaneously to injury probability. The project’s aim is to create an injury predicting algorithm which ranks the different features based on how they affect the risk of injury. The paper also discusses the business potential and strategy of a start-up aiming to provide a solution for predicting injury risk through statistical analysis.

Index Terms—Sports analytics, computer science, machine learning, ice hockey, non-contact injuries, predictive analytics, support vector machine, random forest, SHL

INTRODUCTION

"THe person who develops a solution for predicting injuries will become very wealthy" - Michael Klotz, head of medical team at Hammarby Football [1]. Data analytics and statistical prediction has become an increasingly integrated part of professional sports. Injuries decrease the performance of sports teams, cost considerably in terms of wasted salaries and decrease the marketability of the organization - reducing revenue. Just the salary cost of injuries was estimated to be 218 million USD per year in the NHL 2009-2012 [2]. WinterGreen Research, an independent market research organization, estimates that the sports player tracking and analytics market is expected to grow to 15.5 Billion USD by 2023 [3]. Professional teams have their own analytics departments and large tech companies such as Microsoft and expanding start-ups have developed platforms with the objective to aggregate data and produce key insights through analytics [4]. However, after researching and meeting with several Swedish professional teams only a few seem to perform data analytics at a high level. The focus of current state of the art analytics is almost always directly tied to performance with little to no analysis on injuries [1][5].

In the last decades clubs at the highest level have started collecting more and more data with new electronic measuring equipment in everything from weight lifting accelerometers and GPS trackers to the more common heart rate monitors [6]. Sport clubs in Sweden do perform injury preventing physical training albeit this training is primarily aimed towards improving performance through developing strength and endurance. This training has been tested in a standardized way for some time mostly to monitor improvement and potential weaknesses in strength. Training load has also become subject to monitoring in an attempt to avoid overtraining [7]. What has been lacking so far is the ability to combine and use all the data that has been gathered. It is of great interest for elite clubs to find a useful tool which can predict and give estimations on injury risk [5], which if used correctly could decrease the amount of injuries substantially. Machine learning provides the potential of taking great amounts of data and building a statistical model over many different features, providing a multidimensional approach to the problem. So far, the analysis in most studies and by clubs themselves has been limited to only testing one feature at a time for its correlation with the risk of an injury [5]. This does not capture the reality that different features work together and do not independently affect the risk of injury. The data exists and the clubs are seeking a solution which could provide a competitive advantage. What is lacking to turn ambition into reality is commitment and computational knowledge.

May 25, 2018

REPORT STRUCTURE

This paper is divided into three parts. Part One discusses the implementation and results of a machine learning model for injury prediction in the Swedish Hockey League. Part Two is a qualitative analysis of the market for data analysis for sports clubs in Sweden and the potential business opportunity for a start-up in this market. The two first parts will have their respective research question, theoretical framework, methods, results and discussion. Part Three will be a summative conclusion which ties the previous two parts together.

PURPOSE

The project aims to create a machine learning injury predicting solution based on historical data retrieved from and in cooperation with the Swedish National Register and clubs in the Swedish Hockey League. The purpose of the project is also to evaluate the potential, from a business standpoint, of data analysis for injury prediction. The report will be aimed towards and will be of interest to those working with injury...
prevention, medical teams, physiotherapists, the participating SHL clubs and others that would benefit from a successful injury predictive solution.

Research Questions

How can machine learning be used to predict injury risk in elite hockey? What features are prevalent in injured players and to what extent do these features affect injury risk?

What is the market for an injury predictive solution and how can a solution become a value creating business as a start-up?

Project Scope

In Part One, two models for injury probability prediction, Support Vector Machines and Random Forest Classifier, are built. The models were built to solely predict non-contact injury risk. The data used to create these models is from Djurgården Hockey’s 2013-2018 seasons.

Part Two sets its starting point in Stockholm, Sweden but is not limited to only the Swedish Hockey League, both in terms of sport and region to ensure a broadened analysis.

Part I

A Machine Learning Approach

Theoretical Framework

Previous Studies

Previous sports science research focused on injuries has provided valuable information on different aspects and factors that may affect injury. Factors such as training load, age, ratio between type 1 and type 2 muscle fibers and hamstring strength have all been correlated with the risk of injury [7][8][9][10]. Factors have been assessed individually with results advising athletes to increase muscular strength in different target areas and to keep weekly training load from diverging too much from the seasonal average [11][12].

Research using a multidimensional approach and how multiple factors affect one another is sparse but some example from different sports exist. Two state of the art studies, on basketball in the NBA and on football in Serie B, used machine learning to correlate physical data with injury risk [13][14]. These studies focused on different types of training load with promising results, showing the potential of a multidimensional approach to the issue.

Applied Machine Learning

Machine learning is an emerging technology that has started to become implemented in various types of businesses. Machine learning can be categorized into supervised and unsupervised learning. Supervised learning requires data labelling and parameter tuning in order to improve the quality of the output.

Unsupervised learning uses its own feedback loop and iterates several times to improve. Machine learning consists of several different methods and algorithms, which are specialized in different types of problems such as clustering, classification and regression [15].

K-fold Cross-validation

To test the model, the data set is split into two subsets: one training and one testing data set, where the training set is substantially larger. In order to acquire results which are a representation of the whole dataset, k-fold cross-validation is used. The data set is divided into k mutually exclusive subsets which are then trained and tested on. The outputted model results are then calculated by dividing the combined result from the k-cross-validations by k [16].

Feature Selection in Machine Learning

Feature selection is the process of selecting the features in data that are most relevant for the problem. The result of the selection is a subset of relevant features that are included in the construction of the predictive model. Having fewer attributes is advantageous as it reduces the complexity of the model; a simpler model is easier to understand and explain. Reducing the number of features also reduces the training time and provides enhanced generalization through reduced overfitting.

In essence, feature selection enables an accurate model whilst requiring less data [17]. Feature selection with respect to specific machine learning algorithms will be explored further in coming sections.

Missing Feature Values for Data points

Machine learning has difficulty handling data points with missing values. An approach to the problem is to either remove the data point from the data set or to take the average or the median value from all other data points in that set [15].

Random Forest: Algorithm Description

Random forest is a supervised machine learning algorithm that may be used for both classification and regression problems. For this theoretical discussion, the focus will be on classification. The random forest algorithm operates by constructing several decision trees during training. A single decision tree is a binary tree where each node contains a split condition. How the tree is traversed downwards is thus dependent on the input variable and the split conditions of each node. A single tree provides its own probability of the output belonging to each class as the ratio between the classes in the samples in the leaf the input has landed in. The outputted class is the class with the highest probability after taking the mean over all trees in the forest [20].

The training algorithm for random forest applies the technique of bootstrap aggregation or bagging, where random samples
of training sets are selected with replacement and trees are constructed based on this data. The class prediction for the random forest model is the average prediction of all decision trees constructed in the model. Another central concept of random forest is feature bagging which entails that each node split condition in the decision trees is determined using only a small subset of randomly selected features. The combination of the bagging of data and features diminishes the problem of overfitting commonly associated with decision trees [20].

**Random Forest: Gini Impurity**

As previously described, a splitting criterion is computed at each tree node. There are several methods for computing the feature value which produces the optimal split of data. For classification problems the most common concepts are entropy and Gini impurity. Both methods measure the homogeneity within the split subsets of the target value of the input training data point [20]. Gini impurity will be the focus of this theoretical description. Given a set of labeled data points, the Gini impurity value is the probability of selecting a data point from the set at random and incorrectly predicting the label of the data point, if the label was randomly determined according to the distribution of labels in the set. The Gini impurity for a splitting criterion reaches its minimum (zero) for a binary classification when the splitting criterion divides the two classes into different subtrees [20].

**Random Forest: Hyperparameter Tuning**

Several hyperparameters may be tuned in order to improve the performance of a random forest algorithm. The hyperparameters discussed in this section are: number of trees, maximum number of features and maximum tree depth. How these parameters affect the model is dependent on the amount and characteristics of the data. All three parameters listed below have in common that they negatively affect computation time when increased.

- **Number of trees:** Increasing the number of trees in the forest increases the stability of the model.
- **Maximum number of features:** This parameter determines the number features randomly selected when determining the optimal node splitting criterion. More features generally improve the performance of the model at each node as more options are available, however the diversity of the tree may be negatively impacted if a small set of features are frequently selected.
- **Maximum tree depth:** Increasing the depth of the tree, increases the number of splits which allows the model to capture more information about the data. The risk for overfitting may increase.

**Random Forest: Feature Selection and Feature Importance**

Feature importance may be measured in a random forest algorithm through previously described metrics such as entropy and the Gini index. These metrics describe how effective the features are in dividing data into classes - how important the features are to the model. Feature selection may be performed based on the feature importance ranking to reduce the risk of overtraining.

**Support Vector Machines**

Support vector machines (SVM) are a set of supervised machine learning methods for classification, regression and anomaly detection. The general idea of an SVM is to map all data points in a p-dimensional space, where p is the number of features of the data set. In the case of binary classification, the SVM will try to construct a hyperplane in the p-1 dimension, which separates the data points into two classes. This study will focus on binary classification, but a SVM can be scaled up to more classes if needed. The best separation has the largest distance to the nearest training data point of any class. The points that define this hyperplane are called support vectors [21]. If the data points are not separable with such a hyperplane, the data set is said to be linearly inseparable. That is, if no linear hyperplane can successfully separate the data set into two distinct classes. The solution is to map the data points in higher dimensions through what is known as the kernel trick. This transforms the feature space through a kernel function, allowing the classifier to construct a hyperplane that separates the data points successfully. The radial basis function, polynomial, sigmoid and linear kernel are all examples of common kernels. The choice of kernel is heavily dependent on the training data set and there is no easy way to figure out which one to use. Cross-validation, where the output from each kernel is compared and contrasted, is the best way to choose kernel [22].

**Support Vector Machines: Feature Selection**

A support vector machine with a non-linear kernel does not provide a set of feature importances which can evaluate the weight of each feature. This is because the hyperplane separating the data set exists in a transformed space, different from that of the input space. The weights to each feature in the transformed space are not directly related to the input space; thus, no feature importances can be extracted. However, in the case of a linear kernel the coefficients are in the same space as the input space, since no transformation has been done. As a result, feature importances can be extracted [23].

**Support Vector Machines: Hyperparameter Tuning**

As mentioned previously, the importance of a feature is not regulated through feature selection but rather the adjustment of a hyperparameter called C. The C parameter controls how
penalized a misclassified point should be. A low value of the C parameter penalizes a misclassified point very mildly allowing the hyperplane to have a softer margin at the cost of some misclassification at training. Conversely a high value of C prioritizes minimum misclassification at the cost of overfitting. The gamma parameter defines how far the influence of a class should reach. A low value indicates great influence and high values meaning small influence. The gamma parameter can be seen as the inverse of the radius of influence that selects what support vectors should be used to create the hyperplane. The classifier is very sensitive to the gamma parameter; if the value is too high, the radius of influence only includes the support vectors themselves and leaves no room for regularization with the C parameter and vice versa. Similar to the choice of kernel, the gamma and C parameter can only be selected successfully through cross-validation, where the classifier is fitted multiple times with different values of the hyperparameters as input. The configuration with the best output is chosen.

Unbalanced Classes in Classification

When faced with a scenario where the frequency of classes is unbalanced, some issues arise. Two main methods exist to handle this unbalance. Either the amount of data points from the majority class may be reduced or the amount of data points in the minority class may be increased synthetically. The best solution is to collect more data. Some classifiers, such as support vector machines, random forest and other decision trees, are capable of dealing with unbalanced classes. A classifier trained on a data set with unbalanced classes has a tendency of classifying all data points as the majority class. Thus, when dealing with unbalanced classes, the accuracy metric is not always preferable as a means of evaluating the quality of the classifier.

Probability Behind Classification

In a problem where there is a great unbalance between classes, calculating class probabilities may be a preferable option compared to regular classification. It is possible to calculate a probability of a data point belonging to a specific class in both random forest and SVM, however this calculation is performed in different ways. The machine learning algorithms classify each data point to the class with the highest probability. In random forest a class probability is calculated by taking an average over each trees respective probability. Each trees probability is based on "the fraction of samples of the same class in a leaf". In SVM the probability is calculated using "Platt scaling: logistic regression on the SVM’s scores, fit by an additional cross-validation on the training data".

Method

Consulting Professional Sports Physicians

The first step of the project was meeting with sport clubs. We spoke with the general manager and the medical team of Djurgården Hockey in order to gain insights of the potential for an injury predicting solution based on machine learning. Djurgården has recorded data on training and match load for players and perform standardized physiological tests several times a year. Data on injuries has been recorded in a register, something that has been required for all teams in the Swedish Hockey League for several years. This register was created and is operated by M.D Magnus Forssblad. Meeting with Magnus Forssblad provided information on how to retrieve the data on injuries which is protected by Swedish patient confidentiality laws. A meeting with the head of Hammarby Football medical team, Mikael Klotz, gave insight into how Hammarby as one of the first clubs in Sweden had begun standardizing data collection and performing analysis linked to injuries. Possible features which could potentially correlate to injury in football were discussed. The meeting further proved that there was interest from Swedish clubs for data driven injury prediction however, a collaboration with Hammarby Football was not further pursued for this project as the club lacked sufficient historical data.

Data Retrieval

After being encoded to ensure patient confidentiality, the data was sent to us from both Djurgården Hockey and Swedish National Injury Register. The data received by us includes the training and match load for individual players, the type of training performed, player’s physical profile (several features such as weight and physical tests) and documents describing injuries. A complete set of features may be found in the Appendix. For each injury information was supplied on the type of injury, which player was affected, how the injury occurred and date of occurrence.

Data Preprocessing

Every training or match is an occasion where a player may be injured. Every training or match thus results in several data points, one for each player. The 34,642 data points are classified as either an injury or a non-injury. The injury data points were further classified, based on the comments, as non-contact injury, contact injury or head injury. Some data points on injuries were discarded as they were on youth players for which there was no training load or physical profile for. Some features were dropped as there were not enough data in the data set. The data was further parsed in python to give each data point its corresponding values for all different features. Python was chosen based on familiarity as well as because of the machine learning library scikit-learn also familiar to the project authors.
Model Building in Machine Learning

Different machine learning algorithms were evaluated on what type of algorithm suited the problem. Random forest and support vector machines were chosen based on a number of factors: easy implementation, used in previous similar studies and mitigation of overtraining. Random forest also has a built-in feature importance module in scikit-learn.

As the classes were very unbalanced and the majority of injuries occurred during matches, the training data points were dropped. The injuries that occurred during training sessions were, if deemed reasonable, added to the previous match before that training to ensure that valuable injury data points were not dropped.

The data was then trained and tested in a k-fold split with randomized data points to ensure diversity in the training and testing data. Hyperparameter tuning was performed through cross-validation for both random forest and SVM. For random forest, some features were dropped based on their corresponding importance or rather lack thereof. The feature importances were based on the Gini impurity metric.

Injury Labeling

The predictive model was built to provide injury probabilities for non-contact injuries occurring during matches. This means that all matches that did not result in a non-contact injury were labeled as no injury. Non-contact injuries were selected due to the hypothesis that non-contact injuries would be dependent on training load, match load and the physical profile of players. For concussions or injuries occurring due to physical violence, this correlation was hypothesized to be less evident.

RESULTS

Model Configuration

1) Random forest:
   - Maximum tree depth: 5
   - Number of trees: 100
   - Maximum number of features: 4

2) Support vector machine:
   - Kernel: radial basis function
   - Gamma: 0.001
   - C: 100000
   - Class balance = 'balanced'

Framework for Evaluating Results

All test results were produced and averaged across using a k-fold cross validation, where k=10. Actual injury frequency in the data set was calculated by \( \frac{\text{number of injuries}}{\text{total data points}} \), which the results from the models were compared to. The values produced by the models were calculated on an abstraction level below classification. Instead of binary classification, an injury probability for each data point was calculated. This method was previously explained in the Probability Behind Classification part of the Theoretical Framework.

For each individual data point a probability of the data point being an injury was calculated by the models. For all data points that were an actual injury an average probability was recorded. A higher percentage probability than the actual injury frequency would imply that the model was better at giving a probability of injury on the actual injuries compared to that of guessing. The probabilities for the actual non-injured data points were also averaged over. A lower percentage than the actual injury frequency would imply that the model was better at giving a probability of injury on the actual non-injuries compared to that of guessing. A probability average for actual non-injured data points that is higher than the actual injury frequency would result in an abundance of ‘false alarms’.

Actual Injury Frequency and Model Results

The actual injury frequency on the data set was \( \frac{\text{number of injuries}}{\text{total data points}} \) = 0.97%. This constitutes as a benchmark for each data point in the data set.

Random forest model average probability of injury on actual injury data points: 1.92%. This probability shows that random forest was on average \( \sim 100\% \) better than guessing on the actual injured data points.

Random forest model average probability of injury on actual non-injury data points: 0.81%. This probability shows that random forest was on average \( \sim 15\% \) better than guessing on the actual non-injured data points.

SVM model average probability of injury on actual injury data points: 10.51%. This probability shows that SVM was on average \( \sim 1000\% \) better than guessing on the actual injured data points.

SVM model average probability of injury on actual non-injury data points: 3.94%. This probability shows that SVM was on average \( \sim 400\% \) worse than guessing on the actual non-injured data points.

Lift Curves

The lift curve compares different deciles of the injury probabilities produced by the models, from highest risk of injury to lowest. Using the lift curve, it is possible to see how many true injured data points fall under each decile. For comparison with model results, a diagonal chance line that represents a uniform distribution of injuries across all deciles was plotted. For desirable model results, the lift curve produced by the models would lie over this diagonal line. Lift curves is a way to avoid the misleading values produced by averaging over potential anomalies. Anomalies in lift curves is just one point in a decile and is treated separated from all other points. Using lift curves ensures that the average probabilities presented earlier were not based on skewed injury probability distributions.
Figure 1: Lift curve for random forest displaying percentage of injuries avoided by resting top x % of injury scores.

In Fig. 1 the lift curve for random forest, significantly beats the diagonal chance line. By resting the top 10% of injury probability scores, having the players abstain from playing in the match, 34% of injuries can be prevented. This shows a 240% increase compared to the chance curve. By resting 20% of injury probability scores, 50% of injuries can be prevented.

Figure 2: Lift curve for SVM displaying percentage of injuries avoided by resting top x % of injury scores.

Fig. 2 shows that the lift curve for SVM does not consistently beat the diagonal chance line. Resting players corresponding to a top percentile of SVM injury probability scores would approximately reduce injuries with the same rate as if the players were selected at random.

Random Forest: Feature Importance

Feature importance for random forest is a measure of how effective the features were for classifying injuries and non-injuries. The values sum up to 100%, the higher the value, the more important is the contribution to the prediction function. Feature importance was measured using the Gini metric described in the Random Forest: Gini Impurity part of the Theoretical Framework.

F1: Change in training load on ice from current week compared to this month
F2: Change in training load on ice from current week compared to last week
F3: Accumulated off-ice training load from this month
F4: Accumulated training load on ice from current month
F5: Days since last injury
F6: Accumulated match load from this month
F7: Player weight
F8: Accumulated off-ice training load from this week
F9: Average effect in cycling player weight (physical test)
F10: Power clean one rep maximum (physical test)
F11: Accumulated training load on ice from current week
F12: Change in match load from current week compared to last week
F13: Change in match load from current week compared to this month
F14: Accumulated match load from this week
F15: Squat one rep maximum (physical test)


**DISCUSSION**

**Scrubtny of Results**

Both the random forest and the SVM model provided higher average injury probabilities on actual injury data points compared to the actual injury frequency. However, substantial differences are shown in the average probabilities on the actual non-injured data points. SVM provided a higher probability of injury than the actual injury frequency on the non-injured data points, which if implemented would produce a lot of 'false alarms'. The SVM model's inability to predict injuries is demonstrated clearly in Fig. 2. The high average probability of injury on actual injury data points (10.5%) found by SVM can be explained by a few data points with a very high injury probability and does not mean the model is competent at predicting injuries. If a team would decide to rest the players that had injury probabilities that lied in the top 10% of the model, it would be the case even though the training load did not contain an injury on ice as the most important feature for the model. This was the case even though the training load did not contain an intensity grade. Adding an intensity metric, such as heart rate or distance, to differentiate between low and high intensity training, should theoretically improve the algorithm. For the models based on data from Djurgården Hockey, the physical profile features in general had a lower feature importance than the load features. This could be due to lack of variation in the physical profile data, missing values for physical tests or that the physical profile matters less in terms of injury probability - at least with the high standards of physical strength at Djurgården Hockey. A similar approach for support vector machine with a non-linear kernel is not possible as the model does not provide a set of feature importances. The use of feature importance in random forest shows potential to further deepen the understanding behind what underlying factors affect the risk of injury.

**Sources of Error and Model Improvement**

Two main sources of error and methods to improve the performance of the model have been identified:

1. Adding more data.
2. Building the model on features that are better correlated to injuries compared to the features currently used.

A list of arguments for why more data would be beneficial:

1. The classes were unbalanced.
2. The reliability of the model increases with more data.
3. The model is trained on a limited number of athletes. These players do not necessarily represent the general population of SHL players.
4. Only one club is represented in the data. Training methods and player load may differ between clubs.
5. The clubs represented train differently each season.
6. More data could enable segmentation of players. Players could be segmented upon age or weight. The theory is that models further adjusted to a physical profile would be better at predicting injuries.

Several potential deficiencies in the features used in the models exist. The training load function did not include any kind of intensity or perceived exertion, the training load of a single practice is determined not only by the duration but also by the intensity etc. Another potential improvement would be to compare different load values to the median and/or the variance in a normal distribution of load during a given timespan. There are multiple other features that could affect injury risk that were not included in the predictive model as they were not made available to us by the clubs or had not been recorded. These include age and body fat percentage. Another point of discussion is the strength results, if they should be relative to the athlete’s weight or not.
Part II  
Market Analysis and Strategy for a Start-up

In part Two the potential market is analyzed through the perspective of a start-up formed by the project authors.

THEORETICAL FRAMEWORK

The models and concepts used for this market analysis are:

- SWOT
- Porter’s 5 Forces
- Co-creation
- Customer Education
- First Mover Advantage
- Vertical Integration
- Servitization
- Black Box vs. White Box System

SWOT

SWOT is a strategic analysis tool for identifying the Strengths, Weaknesses, Opportunities, and Threats related to an organization, project or business opportunity. Strengths are internal characteristics leading to a competitive advantage whereas Weaknesses lead to a competitive disadvantage. Opportunities are elements in the external environment that may be exploited to the business or projects advantage. Threats on the other hand are elements in the environment which could jeopardize the business or project.

Porter’s Five Forces

In an article published in the Harvard Business Review, Michael E. Porter published a framework for identifying market competition and how to strategically position a company or business accordingly [29]. The framework was built on five forces, commonly known as Porter’s Five Forces, and it provides a delineation of how to answer the question "What is the potential of this business?". Porter realized that many managers have a very narrow mindset and focuses only on their direct antagonist but fail to recognize the threat imposed by their customer’s and supplier’s bargaining power as well as latent sources of competition. Here follows a description, in no particular order, of what the five forces entail:

1) Threat of new entrants:

New entrants to the market desire to gain market shares and will bring new capacity. If the industry allows for easy entry, fierce competition will follow imminently. Porter therefore proposed multiple sources of barriers to entry, such as economies of scale, capital requirements and product differentiation to name a few.

2) Bargaining power of suppliers:

Powerful suppliers can exert bargaining power on actors in the industry by either raising the prices or lowering quality of their goods and services. A supplier group is considered powerful if, among other factors, its product is unique or differentiated from others and it poses a credible threat of integrating forward into the industry’s business.

3) Bargaining power of suppliers

Customers likewise can force down prices by demanding higher quality for lower price - all at the expense of industry profits. A buyer group is considered powerful if it can, among other factors, purchase in large volumes and the products it purchases are standard and undifferentiated.

4) Threat of substitute products

It is easy to overlook and naively reject the threat of a substitute product, but it can be fatal for any business. Substitute products that from a strategic standpoint require the most attentions are products that improve the price-performance trade-off compared to that of the industry’s.

5) Jockeying for position

Rivalry among existing competitors in the same industry occurs when participants compete by price alternation, aggressive advertising strategy and product introduction.

Co-creation

What adds value in a product for each customer is individual. Even in a business to business setting the receiving company has their separate needs and requests. There is a trend towards co-creation where the customer has an active role during the creation of the product [30]. This is commonly used in software development where iterations with the recipient is especially important [31]. The co-creation aspect of a product improves its overall quality and tailors the product to its target user.

METHOD

Meetings with Djurgården Hockey General Manager, Joakim Eriksson, and Other Potential Stakeholders

To survey the interest for a potential injury predicting solution, several meetings were held with key stakeholders in Swedish top-league clubs in several sports. Information was gathered on their potential interest, what data they had to offer, what other kind of solutions they are working with currently and how a solution would help their business.

Market Estimation

Analysis was based on public sources of information and insights provided by Joakim Eriksson, General Manager of Djurgården.
Applying SWOT and Porter’s 5 Forces

These models allow for a deepened understanding of the potential start-up’s role and the market it would act in.

RESULTS

SWOT

1) Strengths

1) Connection established with Djurgården IF Ice Hockey, Frölunda Hockey and HV71s.
2) Connection established with Magnus Forssblad responsible for the Swedish National Injury Register where all SHL clubs keep their data.
3) Existing framework for adding data from other teams to prediction model.
4) Network among university students who could work on the project.
5) Developing knowledge in machine learning.
6) Team members have background in elite sports.

2) Weaknesses

1) Difficult to explain machine learning to those working in elite clubs.
2) Difficult to acquire sufficiently detailed data quickly.
3) Sensitive injury data requires a middle man to encode the data.
4) Difficult to convey legitimacy for complex and intangible products.

3) Opportunities

1) Expansion to several different sports.
2) Large potential market.
3) Great interest among sport clubs.
4) Several big elite clubs in the local area.
5) Potential tangible product to market - platform for data collection.
6) Funding from KTH start-up incubators and potentially sport clubs.
7) Vertical Integration. Potential associated products - GPS equipment, physiotherapy equipment etc.

4) Threats

1) Other partnership already in place.
2) Bigger actors with already existing platforms.
3) Clubs selling their exclusive right to the data.
4) Skepticism towards data driven analysis among authoritative figures in sports.

Market Estimations

Several independent organizations have estimated a rapid growth in the data analytics market in sport in the coming years [32]. The market has been estimated to grow up to 15.5 billion USD worldwide by 2023 with no indication of stopping there [3]. Apart from companies focused on data analytics, the market also consists of tangible products such as GPS tracking devices and sensors. As the products become better and more widespread more data is stored and can be analyzed. The two parts of the market, product and analytics go together hand-in-hand. The major team sports which make up a great part of the market are American football, football, hockey, basketball and baseball [32].

The Swedish Hockey League consists of 14 teams and has a total revenue of 1700 million SEK 2017, a number that is expected to increase in coming years due to a new television deal [33]. SHL’s revenue although impressive is around 20 times less than the National Hockey League [34]. Personnel costs make up 61% of all costs for clubs in the SHL [33] often with extensive medical teams to deal with injuries amongst other issues. What can be concluded is that the market for a successful solution is large, even in the more regional areas in and around Sweden.

Takeaways from Meeting with Djurgården General Manager

• Need for software platform for data management, medical records, schedules
• Interest for statistical analysis within organization
• Have been approached by companies providing statistical analysis for on ice performance but not by companies providing a data-driven injury prevention
• Willingness to purchase and use intensity measuring equipment such as heart rate monitors
• Willingness to invite data analyst to evaluate data gathering practices during pre-season camps
• Willingness to cooperate with other clubs
• Initial budget of around 50-150 000 SEK for data analysis depending on the service [5]

Porter’s 5 Forces

1) Jockeying for position

Several companies, such as Kitman Labs, KINDUCT and EDGE10, have built businesses that include injury data analysis for professional sports teams. These companies all have player management platforms where clubs can record information and data on players, communicate internally and sync schedules. While all the companies use formulations like "we provide key insights through big data" [35] the degree of analysis performed is difficult to determine. Large tech companies like Microsoft or Ericsson have performed projects to apply machine learning and predictive analytics, although these projects have been experimental partnerships with single sports clubs and not part of a serious business venture [36][37].

2) Threat of new entrants

Although companies providing data-driven injury analysis have some of the largest clubs in the world as clients, many professional sports clubs do not employ such services. There
is thus considerable potential for new entrants to enter the market whether they be small start-ups or large existing tech companies. A more established industry connected to elite sports is companies that specialize in wearables, including heart rate and GPS monitors. Companies such as Zephyr Technology Corp, Zebra Technologies International, Fitbit and Catapult today primarily provide the hardware and let clubs analyze the data themselves [38]. It is far from inconceivable that these companies envision a vertical integration where they instead perform the analysis for the clubs. A barrier of entry to deter the threat of new entrants is first mover advantage and economies of scale. Already having a functional platform for recording and visualizing data and having access to large data sets from several clubs are both significant competitive advantages.

3) Bargaining power of suppliers
The most important resource provided by suppliers is data. An interesting dynamic occurs as the clubs are both the suppliers of data as well as the buyers of the analysis service. It should thus be in the clubs’ best interest to sufficiently supply historical data or record current data in order to fully reap the benefit of the resources invested in the product. If historic data is not owned by the clubs but instead by a third party e.g. by the league, an unwillingness may exist to supply historical data in an efficient manner. As presumably the sole owner of this data, the third party exhibits a great bargaining power. The data analysis may be seen as a threat to this third party or there may simply not be a clear incentive.

4) Bargaining power of buyers
The bargaining power of customers is tied to how standardized the service is and the scale of the customer as well as the organization providing the solution. Creating a platform and achieving scale across several teams and leagues decreases the bargaining power of customers compared to working on an individual consultant basis using historical data. The latter example where analysis is performed on more of an individual basis at the customers discretion, decreases predictability which could affect margins.

5) Threat of substitute products
Several other products and approaches, such as training or medical equipment, exist to decrease the risk of injury. While these approaches can be viewed as compliments and not a substitute for data analysis in the context of a club’s budget with limited resources, priorities have to be made [1].

**Weakness 1:** Those working with injuries in elite clubs have a lot of experience and are specialized in other methods which are not statistically driven. Machine learning is unfamiliar and requires some initial knowledge in both statistics and computer science to understand. Thus, educating he clients in machine learning and what benefits it may provide in terms of decreasing injuries requires personal meetings.

**Opportunity 1:** Similar to strength 3, a framework can be adapted other sports providing a whole new market. Different sports have different types of data features but share the issue of injuries and the opportunity of machine learning.

**Opportunity 2:** Sports player tracking and analytics market is expected to grow to 15.5 Billion USD by 2023 [3].

**Threat 4:** Historically the attitude in sport has been somewhat skeptical to data-driven analysis. There is a culture valuing experience and gut feeling which can be difficult to sway.

**Co-creation**
Co-creation entails that the perceived value is increased when the customer, in this case sport teams, take an active role in the development of the product. Data-driven injury prediction as a solution is characterized by themes suitable for co-creation. The customer needs to record or provide the data which the algorithm is based upon, which increases its performance. This gives an incentive for the clubs to be more engaged in the development of the product, which provides a sense of ownership of the product.

**Black-box vs. White-box**
Using a black-box approach, meaning that an injury probability is provided without any explanation for the underlying factors, presents many issues for real applications by clubs. Djurgården Hockey and Hammarby Football explicitly stated that "magical solutions" have been promised before with unexplainable methods which made the clubs deem the services unserious and not evidence based. Thus, a potential injury predictor needs to show the customer what affects the risk of injury and also preferably just how much. If a player is at risk of injury, clubs need to know what measures to take to mitigate this risk e.g. does the player need to adjust the amount or intensity of training or increase his physical capabilities.

**Injury Prediction as a Service and Customer Education**
A condition for clubs gaining an insight into the factors which affect the prediction model is a continuous interaction with the customer and a view on the product as a service rather than a one-time sell. Building a relationship with the customer is especially important as the customer value increases over time with added data and familiarity with the service. Educating the customer of the opportunity that a machine learning prediction model presents is not only necessary in an initial selling stage.
Continuously working with clubs educating them on statistics, presenting data on players in a comprehensive manner and tying findings to traditional sports injury research will have a positive effect on customer trust and loyalty. Another concrete measure in such an approach is the clubs having a personal contact with knowledge in both sports medicine, athletic training and data analysis for weekly follow-ups with the clubs’ medical team. Having people within the organization with knowledge of traditional sports medicine and training is essential in order to be able assist the clubs with fully implementing the product in everyday operations. The benefit of a customer relationship built on interaction and education when providing a technologically complex and intangible product is summarized in a study on customer education by Eisingerich et al: “Faced with highly complex and intangible service products, customers perceive an organization’s effort to provide essential information as an important and valuable service augmentation” [39].

Developing a Platform

Building a platform where clubs can record player data has several advantages over working with historical data. The obvious benefit is that not all clubs have sufficient historical data on e.g. player training load, physical test and injuries in order to build a model for injury prediction, this greatly limits the number of potential customers. Creating a platform standardizes what data and features that are recorded across clubs, leagues and sports. This aggregation of data is a huge benefit, perhaps even a prerequisite, for statistical modelling. Analysis over larger more diverse data set allows more accurate, reliable results and identification of detailed trends. Aggregation of data thus presents an economies of scale advantage. Building a platform would significantly decrease the marginal cost of adding new clubs as customers. Steps such as the reviewing and formatting of historical data would be unnecessary if current data is used. The decreased need for initial customization does however not diminish the importance of the regular customer interaction discussed in previous sections. A platform also gives the opportunity to provide functions that could give more of an overall solution for player tracking. Such possible functions include support for medical records, recordings of physical tests and player training loads, appealing graphs displaying players development or support for GPS trackers.

The drawbacks of developing a platform is that relying on current data requires a start-up period before sufficient data for analysis has been collected. This start-up period can be reduced considerably if scale is achieved across sports and leagues is achieved. Although all players and teams are different, it is probable to assume that many similarities exist for teams playing in the same league. There is naturally also a great initial cost associated with developing a platform; a substantial risk is involved.

Discussion on Takeaways from Meeting with Djurgården General Manager

After meeting with Djurgården general manager it became clear that to develop and launch a successful product for SHL and Djurgården, what is needed is cooperation with several clubs and incorporating services other than injury prediction, such as data visualization and administrative software. The advice of starting with several clubs in order to fund the project was given, otherwise external investors or capital needs to be involved. The budget for a statistical service not limited to only evaluating data on injuries would initially be around 50-150 000 SEK. Budget items such as player and staff salary are more heavily prioritized. Thus, to establish and create a project several clubs need to be brought on board. There is willingness to cooperate in terms of purchasing heart rate monitoring equipment and using it during training thus adding necessary detail and improving the possibility of a successful model. Furthermore, Djurgården is willing to cooperate with other clubs and not having exclusive rights to the service. They were also open to evaluate and change their data gathering practices to assist the performance of statistical solution [5]. Djurgården is a big club in the SHL and could potentially attract other key clubs to the project. Their interest and established relationship provides legitimacy that is vital for the project. Djurgården is a great starting point but other clubs might be needed to provide the funds and amount of data necessary.

Part III

Conclusion

A key takeaway from the machine learning models’ results of this thesis was the need for more data. This leads to a couple of interesting implications from a business perspective. It raises the question of the need for initial scale. For a potential start-up this means having several clubs involved before releasing any product or service. Furthermore, it brings up the question, which was discussed in part two, of creating a platform where clubs could record data. This would enable control over data collection and the features included in a predictive model. A platform could also provide additional value to clubs through support for administrative tools such as medical journals, scheduling and data management.

The issue with creating a platform, even if its desirable in the long term, is that it requires time and capital. Time is of essence in order to be first on the market with this new developing technology. If dealt with strategically it is more logical to expand as much as possible, developing relationships with several different SHL clubs making them part of the process, before developing a platform. If successful a product can be developed with sufficient data and can be piloted with involved clubs to then together with the customer/clubs develop a platform and expand further. After that, horizontal integration into different sports and countries would be possible.

In this paper an attempt was made to use the machine learning algorithms of random forest and support vector machines to
predict injuries in elite hockey. The lack of sufficient amounts of data and the features available in historical data proved critical. The features prevalent in injured players, at least in this paper’s scope, is still speculative with different importance to injury risk. The market for an injury predictive solution based on data analysis such as machine learning has a lot of potential.

REFERENCES

[1] Personal meeting, 20-02-18, Stockholm Odenplan, Michael Klotz, Head of Hammarby Football medical team
[7] Lüks et al. The preventive effect of the nordic hamstring exercise on hamstring injuries in amateur soccer players: a randomized controlled trial 2015
[30] NHL, December 8 2017, NHL revenues, salary cap projected to rise, NHL.com, Dan Rosen senior writer
APPENDIX

Complete Set of Features Available for Models

1) Accumulated training load on ice from current week
2) Accumulated training load on ice from current month
3) Accumulated off-ice training load from current week
4) Accumulated off-ice training load from current month
5) Accumulated match load from current week
6) Accumulated match load from current month
7) Change in training load on ice from current week compared to last week
8) Change in training load on ice from current month compared to last month
9) Change in off-ice training load on ice from current week compared to last week
10) Change in off-ice training load on ice from current month compared to last month
11) Change in match load from current week compared to last week
12) Change in match load from current month compared to last month
13) Standard deviation of training load on ice from current week compared to last week
14) Standard deviation of training load on ice from current month compared to last month
15) Standard deviation of off-ice training load on ice from current week compared to last week
16) Standard deviation of off-ice training load on ice from current month compared to last month
17) Standard deviation of match load from current week compared to last week
18) Standard deviation of match load from current month compared to last
19) Number of days since the player last suffered an injury
20) Number of injuries the player has suffered within a year
21) Player Weight
22) VO2 max (physical test)
23) l/min (physical test)
24) Time bike (physical test)
25) RER time (physical test)
26) Max pulse (physical test)
27) 0-5 w/kg (physical test)
28) PP w/kg (physical test)
29) Mean w/kg: Average effect in \( \frac{\text{cycling}}{\text{playerweight}} \), (physical test)
30) Min w/kg (physical test)
31) 25-30 w/kg (physical test)
32) Perc drop (physical test)
33) Clean one rep maximum (physical test)
34) Squat one rep maximum (physical test)
35) Bench one rep maximum (physical test)

The physical tests are standardized, although some tests are specific to Djurgården Hockey. For questions regarding the physical test please contact Christian Schumacher, Physiotherapist at Djurgården Hockey.

Author presentation: The authors of this bachelor thesis are students of Industrial Engineering and Management with technological specialisation in Computer Science and Communication at the Royal Institute of Technology. The authors have a background in elite sports, specifically ice hockey, football and basketball, both as players and youth coaches.

SAMMANFATTNING
