CCT College Dublin

Assessment Cover Page

Module Title:	Strategic Thinking
Assessment Title:	CA 1 - Capstone Project Proposal
Lecturer Name:	James Garza
Student Full Name:	Kavi Patak
Student Number:	sba22391
Assessment Due Date:	29 October 2023
Date of Submission:	29 October 2023

Kavi Patak

Declaration

By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution.

Leveraging Machine Learning and Data Science for Competitive Advantage:

Estimating Player Market Values

by Kavi Patak

sba22391

Strategic Thinking (M1)

Lecturer: James Garza

CCT College, Dublin

October 29, 2023

Table of Contents

Introduction

Organisations have historically implemented enterprise systems such as enterprise resource planning in attempts to gain a competitive advantage (Goundar, 2021). The advent of the internet and proceeding developments in technologies like the Internet of Things(IoT), cloud computing, block chain, big data, Machine Learning (ML) and Artificial Intelligence (AI), have heavily influenced enterprises systems of today, and opened new avenues for conducting business.

Organisations have realised the potential value that resides in data and thus search for ways of utilising this valuable asset, it is here that Business Intelligence (BI) has become an important concept (Agarwal & Dhar 2014, cited in Persson and Sjöö, 2017). BI is an organisation's ability to effectively use the information it collects from daily enterprise. (Vidal-García et al., cited in Niu et al., 2021). By identifying emerging opportunities, highlighting potential risks, providing useful insights and supporting decision making, ensure BI plays a significant role in optimising organisational effectiveness (Zhao et al., cited in Niu et al., 2021).

The role of analytics in football has evolved over the past decade, and will continue to do so. Technological developments continue to improve the volume and quality of data available to the world's leading clubs, as well as the ability to derive insight from them. The opportunity exists for clubs of all sizes to use analytics to build a sustainable competitive advantage, something which will be most evident in the area where they invest most: the transfer market.

Having previously assessed the potential for a club to gain a competitive advantage through AI by means of Porter's Five Force Framework (Patak, 2023), this research will focus on leveraging machine learning and data analytics in player scouting and recruitment. More precisely, this report will outline the proposed steps in developing a football player value assessment model using machine learning techniques and in doing so, aid a football club in making more objective and data driven transfer decisions.

This research project will follow the Cross-Industry Standard Process for Data Mining (CRISP-DM) approach in iterating through the stages of Business Understanding, Data Understanding, Modelling, Evaluation and Deployment.

Problem Domain and Objectives

The first phase of the CRISP-DM lifecycle calls for an understanding of the domain and business objectives, and extracting from this the requirements and goals for this project (Wirth and Hipp, 2000).

Football is a highly lucrative sport which depends heavily on the services of football players, the main suppliers to a football club (Patak, 2023). It is therefore no surprise that the majority of a club's expenditure is on player transfers and wages, with smaller clubs unable to compete financially with their larger counterparts (Metelski, 2021). Furthermore, the consequences of poor scouting and recruitment can be devastating to a club, regardless of size and can have adverse affects not only from a business perspective (Depken and Globan, 2020), but to squad harmony, on pitch result and even to a clubs reputation.

This project aims to level the playing field, at least financially in the transfer market, by developing a machine learning model to estimate the market value of football players based on various features such as age, goals, assists and other contributing factors.

Hypothesis: A fair and accurate player market value can be estimated through ML from the available features within the selected dataset.

In addition to this core objective, this project aims to provide insights into promising players who may be undervalued in the market. In-depth analyses shall be conducted on player performance data in specific positions, based on certain physical attributes, as well as exploring the relationship between a players age, value and performance.
Finally, this project will compare the market values of different teams and investigate any changes in player market values over time.

Scope and Methodology

This capstone project will span over two semesters and involve in-depth analysis and exploration of football players and game statistics in developing a football player value assessment model.

Having developed a business understanding of the task at hand, and following the CRISP-DM methodology, the remaining steps in the development lifecycle call for Data Understanding, Modelling, Evaluation and Deployment.
The CRISP_DM methodology is “Agile” in nature, and unlike the traditional linear Waterfall lifecycle, the sequence of phases is not strict (Wirth and Hipp, 2000). In identifying a viable use case for this capstone project, both the Business Understanding and Data Understanding stages have been initiated as one is intrinsically linked to the other. A brief exploration of the datasets has been conducted to confirm that it will satisfy its need.

In adhering to the CRISP_DM lifecycle, the following processes will be carried out:

- Consolidating and Cleaning the Data,
- Exploratory Data Analysis (EDA),
- Feature Engineering,
- Model Selection,
- Training and Validation,
- Defining Evaluation Metrics,
- Tuning Hyperparameters,
- Predictive Analytics,
- Producing Visualisations,
- Further iterations and improvements.

Consolidating and Cleaning

The selected dataset is called “Football Data from Transfermarkt” provided by Kaggle and available at:
https://www.kaggle.com/datasets/davidcariboo/player-scores/versions/284

This dataset consists of nine CSV files with information on football competitions, games, clubs, players and player appearances. Each file contains attributes of the entity and the ID's that will be used to join them. The collective dataset consists of 124 columns (features) and over a million rows (observations) in certain files.
The datasets will need to be further explored, consolidated and cleaned by handling any missing values, outliers and inconsistencies.

Exploratory Data Analysis (EDA)

EDA will be conducted using statistical and visual techniques to better understand the structure and relationships within the data. EDA is a process of examining the available dataset to discover patterns and trends and to identify correlations and interesting insights within the data. It will aid in spotting anomalies, testing hypotheses and clarifying any assumptions (Suresh Kumar Mukhiya and Ahmed, 2020).

Feature Engineering

Feature engineering is an invaluable process in developing and enriching machine learning models. It includes feature generation, feature extraction and feature selection using open source tools such as Jupyter Notebook, NumPy, SciPy, pandas, and scikit-learn libraries. Feature generation entails creating new features from the existing features within the dataset. These new features are engineered to provide additional information that may be relevant and useful for the problem at hand. For example, a player's age may be deduced from their date of birth. Feature extraction is the process of reducing the size or dimensionality of a large dataset while feature selection involves choosing a subset of the most relevant features by removing redundant or irrelevant rows that may introduce noise or lead to overfitting (Rahul Kumar, 2019). Random forest importance and correlation coefficient methods will be considered.

Model Selection

Model selection is choosing the appropriate machine learning algorithms based on the objectives at hand. The objective is to predict the market value of players which is continuous data, making this a regression problem. Alternatively, the dependent or target variable can be categorised and assigned a discrete numerical value for testing classification models. The following supervised models in logistic, linear and multiple linear regression, decision trees, random forest and KNN will all be explored with nothing definitively ruled out at this early stage of the project.

Training and Validation

This step requires splitting the dataset into training and validation sets. The training set which allows the model to learn and the validation set on which we test how well the model has learned. Multiple instances of different training to validation set ratios may be tested including 70:30, 80:20 and 90:10 while studying the training error, validation error and model accuracy. It is not uncommon for ML practitioners to split the dataset into three subsets, one each for training, validation/development and testing. The development set can be used to assess different models performance, tune parameters and minimise overfitting (Rahul Kumar, 2019).
In selecting and training a model it is important to be aware of overfitting, underfitting, bias and variance. To avoid overfitting, regularisation techniques such as Ridge Regression (L2 Norm) or Lasso (L1 Norm) can be implemented. Additionally, cross-validation using K-fold cross-validation will be used in validating and evaluating each model to identify which model is performing better.

Evaluation Metrics

Here we will define metrics to evaluate the performance of each model. Apart from cross-validating each model, a confusion matrix can be used for true positives, true negatives, false positives and false negatives for any classification models to visualise the results in matrix form. Other important metrics that will be measured include accuracy, precision, recall and F1 score which is the harmonic mean of recall and precision. Sklean also provides mean absolute error, mean squared error and R2 for more advanced metrics, with R2 being more intuitive in evaluating regression models (Müller and Guido, 2017).

Tune Hyperparameters

Hyper parameter tuning is the process of selecting the optimal set of hyperparameters for a machine learning model. This is an important step in the development of a model as the choice of parameters can have a significant impact on performance. Manual, Grid, Random and Bayesian Search methods will be considered.

Predictive Analytics

Once satisfied with the training and performance of a model, it can be used to make predictions on new data.

Visualisation

Visualisations and graphs will be used throughout the development process in EDA, feature engineering, modelling and finally to communicate the research findings effectively. Open source libraries like matplotlib and seaborn in Python will be used in a Jupyter Notebook in developing and documenting the process.

Continual Improvement

Multiple iterations of the CRISP-DM lifecycle may be necessary to reassess and refine the model based on new data or insight, and to ensure that the project remains on track in achieving its objectives.

Boundaries and Limitations

It is important to note that each club is unique and so to their structure, ambitions and vision. What is right for one club may not be for another. Some aspire to win the league while others to avoid relegation. Other business models may depend solely on selling players for profit (Sloane, cited in Van den Berg, 2011). Hence aligning and framing scouting requirements with the vision of a club is crucial, as is the recognition that talent id is one small part of squad evolution and building. Another challenge may reside in harnessing the data and bringing together different data sources. There are multiple factors that contribute to evaluating a player's value, some of which may not be available within the selected dataset. Aspects such as exchange rates and social or economical factors like war, recession or a worldwide pandemic may prove challenging to account for (Metelski, 2021).

Timeline

Semester One:

- Data preprocessing and cleaning: 6 weeks

- Feature engineering: 6 weeks
- Model selection: 5 weeks
- Training and validation: 6 weeks

Semester Two:

- Tuning hyperparameters and defining evaluation metrics: 4 weeks
- Predictive analytics and model testing: 4 weeks
- Producing visualisations: 5 weeks
- Documentation and reporting: 3 weeks

This is a very high level timeline with approximate estimates on time that ensures a phased approach to the project.

Data Sources and Ethical Considerations

The development and deployment of ML models raise ethical questions. Issues such as privacy, data security, bias in algorithms, fairness, and the potential for AI to discriminate or negatively affect vulnerable populations in society requires careful consideration.

The chosen dataset is called “Football Data from Transfermarkt” and is provided by Kaggle, available at: https://www.kaggle.com/datasets/davidcariboo/player-scores/versions/284
This dataset consists of nine CSV files with real information on football competitions, games, clubs, players and player appearances. The collective dataset consists of 124 columns (features or variables) and over a million rows (observations) in certain files with varying data types. This large dataset has been selected to allow for sufficient scope and challenges in data preprocessing and cleaning, feature engineering and for in-depth analysis and exploration spanning over two semesters.

It is an open source or publicly available dataset that has been granted a CC0 dedication which acts as a “licence from the affirmer granting the public an unconditional, irrevocable, non exclusive, royalty free licence to use the work for any purpose”, “one can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission” (wiki.creativecommons.org, n.d.). Copyright laws differ by region, and an automatic copyright may be granted in certain jurisdictions. By using CC0, one signals to the public that they relinquish any such rights.

Tracing back the source of the data even further we find that it has been scraped from the TransferMarket website. Data scraping is now a common practice but GDPR requires controllers to adhere to its key principles in purpose limitation and data minimisation. According to fieldfisher website, “purpose limitation means that businesses should only collect and process personal data to achieve specified, explicit and legitimate purposes and not engage in further processing unless it is compatible with the original purpose for which the data was scraped” and “data minimisation means that businesses must only collect and process personal data that is relevant, necessary and adequate to accomplish the purpose for which the data was scraped”, (Campbell, 2019). Moreover, not all personal data can be scraped. GDPR regards certain categories of personal data such as race, religion, health data and political opinions as special categories requiring extra levels of protection (Campbell, 2019).

Pannucci and Wilkins believe that “bias can occur in the planning, data collection, analysis, and publication phases of research”. Following the classic “garbage-in-garbage-out” expression, ML algorithms can perpetuate biases through the use of big data that reflects these biases (Solon Barocas and Selbst, 2015). While it is difficult to completely eliminate bias, careful consideration will be given to the features used in this ML project in attempts to avoid introducing bias. Furthermore, efforts will be made to ensure fair representation of players from various leagues, clubs, and positions. Evaluation metrics will be used in developing the model to help account for fairness and to improve accuracy. This will aid in building a model that generalises well to diverse scenarios. Finally, it is important to demonstrate transparency in the development of the model and in communicating the results and findings, including any limitations or uncertainties associated with the project.

Conclusion

Player scouting and market value analysis are key areas where ML can provide valuable knowledge and insights to the football industry. This capstone project aims to contribute to this by empowering clubs and stakeholders with data driven decision making capabilities for a competitive advantage. By leveraging machine learning on the “Football Data from Transfermarkt” dataset, the project aligns with the evolving landscape of sports analytics, providing practical solutions to real world challenges in the football industry.

References

Agnellutti, C. (2014) Big Data: an Exploration of Opportunities, Values, and Privacy Issues. New York: Nova Science Publishers, Inc (Internet Theory, Technology and Applications). Available at: https://search.ebscohost.com/login.aspx?direct=true&db=e020mww&AN=811106&site=eds-live

Campbell, F. (2019). Data Scraping – Considering the Privacy Issues. [online] Fieldfisher. Available at: https://www.fieldfisher.com/en/services/privacy-security-and-information/privacy-security-and-information-law-blog/data-scraping-considering-the-privacy-issues.

Depken, C.A. and Globan, T. (2020). Football Transfer-Fee Premiums and Europe’s Big Five: Online Appendix. SSRN Electronic Journal. doi:https://doi.org/10.2139/ssrn.3617260.

Goundar, S. (2021). Enterprise systems and technological convergence : research and practice. [online] Charlotte, NC: https://eds.p.ebscohost.com/eds/ebookviewer/ebook/ZTAyMG13d19fMjczMDYwNF9fQU41?sid=997bf805-38e3-47f1-bfb3-813e10171b3f@redis&vid=3&format=EB

Metelski, A. (2021). Factors affecting the value of football players in the transfer market. Journal of Physical Education and Sport, 21(2). doi:https://doi.org/10.7752/jpes.2021.s2145.

Müller, A.C. and Guido, S. (2017). Introduction to machine learning with Python : a guide for data scientists. Beijing: O’reilly.

Niu, Y., Ying, L., Yang, J., Bao, M. and Sivaparthipan, C.B. (2021). Organisational business intelligence and decisionmaking using big data analytics. Information Processing & Management, [online] 58(6), p.102725. doi:https://doi.org/10.1016/j.ipm.2021.102725.

Pannucci, C.J. and Wilkins, E.G. (2011). Identifying and Avoiding Bias in Research. Plastic and Reconstructive Surgery, [online] 126(2), pp.619–625. doi:https://doi.org/10.1097/prs.0b013e3181de24bc.

Patak, K. (2023). Strategic Analysis of Emerging Technology for Competitive Advantage: Artificial Intelligence in the Football Industry. [online]
Available at: https://drive.google.com/file/d/1g7y_U2cSYDq771WJZoka4AvgIkdOvIW7/view?usp=sharing

Persson, J. and Sjöö, E. (2017). Business Intelligence - its impact on the decision making process at higher education institutions .

Rahul Kumar (2019). Machine learning quick reference : quick and essential machine learning hacks for training smart data models. Packt Uuuu-Uuuu.
Solon Barocas and Selbst, A.D. (2015). Big data's disparate impact. Ssrn Elibrary.

Van den Berg, E. (2011). The Valuation of Human Capital in the Football Player Transfer Market. [online]
Available at: Link

Suresh Kumar Mukhiya and Ahmed, U. (2020). Hands-on exploratory data analysis with Python : perform EDA techniques to understand, summarize, and investigate your data smartly. Birmingham: Packt Publishing.
wiki.creativecommons.org. (n.d.). CC0 FAQ - Creative Commons. [online]
Available at: https://wiki.creativecommons.org/wiki/CC0_FAQ.

Wirth, R. and Hipp, J. (2000). CRISP-DM: Towards a Standard Process Model for Data Mining . Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, (Vol. 1, pp. 29-39).

World Population Review (2022). Most Popular Sport by Country 2020. [online] worldpopulationreview.com.
Available at: https://worldpopulationreview.com/country-rankings/most-popular-sport-by-country

www.kaggle.com. (n.d.). Football Data from Transfermarkt. [online]
Available at: https://www.kaggle.com/datasets/davidcariboo/player-scores/versions/284/data