Project C20: Most Expensive Footballers 2021

Team:
Kirill Ryrmak

Task2. Business understanding

1) Identifying my business goals:

I would say that the management’s problem is that the database needs to be constantly updated, to keep track of the number of players and their special characteristics, taking into account their level of play, injuries, frequent transfer price fluctuations and so on. At the end of each season, you will need to completely update the database, making an incredible amount of changes in the characteristics of players and their rating. The goals of the business are to create an optimal ratio of player ratings based on up-to-date information and not including minor errors in the database.

The limitations of this work will be such that it will not be possible to make reasonable low-level football players who are not included in the top 500 football players at the moment, but only using fresh statistics and a high rating.

This problem may not fully reflect the correctness of the data, since it will be necessary to constantly monitor the development of the player, his qualities, the transfer price and which team he plays for. To do this, it will be necessary to make decisions and create different methods for predicting and processing new data.

Background:

Our client wants to select an average Player of the Month by extracting the average salary, price, club popularity, and nationality from a table. Thanks to this data, the client can easily get information about this player. The database allows new player data to be used for personal comparison, consistency statistics of one player as the best, and to create new player icons for football-related games.

The commission is going to check the accuracy of the data based on their government connections and a detailed forecast. Since the data is updated every season and not every day, there will be a discrepancy. Accordingly, it will be necessary to provide an annual report on summing up the formed list of top players, making changes to all features.

Earlier research may include adding important metrics like player performance in each match, training success, etc. Therefore, it will be necessary to complete the models to compile a list.

Business goals:

The goals of the business, in my opinion, will be to increase the number of top players by 3 times, which will lead to more successful tracking of players below the top 500 and will be able to perform statistical calculations with a huge number of players. Also, the purpose of the business may be the role of a scout, that is, the method of distribution is to try to determine the football club for the player himself, based on his data.

Business success criteria:

It is planned to set a specific task for execution, but there are no special evaluation criteria. There will only be a Completed / Not Completed label, and the deadline for this period.

2) Assessing your situation:

Inventory of resources:

Absolutely, the business will contain an owner who is responsible for promoting his business, workers who will have different goals and tasks depending on mentors, those support to help customers navigate the product, the main server that stores information regarding databases, their implementation changes made, etc.

Requirements, assumptions, and constraints:

The business will have a legal agreement to use specific databases, the security will be an officially signed guarantor and customer testimonials.

Risks and contingencies:

It seems to me that the only problem will be network outages when working remotely, since one way out of this situation will be virtual machines, but without saving changes, only in scope.

Terminology:

Accuracy

Accuracy is an important factor in assessing the success of data mining. When applied to data, accuracy refers to the rate of correct values in the data. When applied to models, accuracy refers to the degree of fit between the model and the data. This measures how error-free the model's predictions are. Since accuracy does not include cost information, it is possible for a less accurate model to be more cost-effective. Also see precision.

API

An application program interface. When a software system features an API, it provides a means by which programs written outside of the system can interface with the system to perform additional functions. For example, a data mining software system may have an API which permits user-written programs to perform such tasks as extract data, perform additional statistical analysis, create specialized charts, generate a model, or make a prediction from a model.

Decision tree

A tree-like way of representing a collection of hierarchical rules that lead to a class or value.

Layer

Nodes in a neural net are usually grouped into layers, with each layer described as input, output or hidden. There are as many input nodes as there are input (independent) variables and as many output nodes as there are output (dependent) variables. Typically, there are one or two hidden layers.

Mean
The arithmetic average value of a collection of numeric data.

Median
The value in the middle of a collection of ordered data. In other words, the value with the same number of items above and below it.

Missing data
Data values can be missing because they were not measured, not answered, were unknown or were lost. Data mining methods vary in the way they treat missing values. Typically, they ignore the missing values, or omit any records containing missing values, or replace missing values with the mode or mean, or infer missing values from existing values.

Mode
The most common value in a data set. If more than one value occurs the same number of times, the data is multi-modal.

Precision
The precision of an estimate of a parameter in a model is a measure of how variable the estimate would be over other similar data sets. A very precise estimate would be one that did not vary much over different data sets. Precision does not measure accuracy. Accuracy is a measure of how close the estimate is to the real value of the parameter. Accuracy is measured by the average distance over different data sets of the estimate from the real value. Estimates can be accurate but not precise, or precise but not accurate. A precise but inaccurate estimate is usually biased, with the bias equal to the average distance from the real value of the parameter.

Regression tree
A decision tree that predicts values of continuous variables.

Training
Another term for estimating a model’s parameters based on the data set at hand.

Costs and benefits:

The project will not include special benefits, but in terms of earnings, if our company recreates the necessary and effective methods for solving existing problems, then the rate will increase for each problem.

3) Defining my data-mining:

Data-mining goals:

Based on the purpose of the business, I will incrementally add players weekly to meet the business conditions and create a quality or average player prediction model with a new data stream and maybe to propose new dataset with other non-top players.

Data-mining success criteria:

The model accuracy will reflect the precision of done work of the members but all time it needs development cause the adding new data to the existing dataset will destroy previous model accuracy.

Task 3. Data understanding

1) Gathering data:

Outline data requirements:

Predictive Data Mining

As the name signifies, Predictive Data-Mining analysis works on the data that may help to know what may happen later (or in the future) in business. Predictive Data-Mining can also be further divided into four types that are listed below:

- Classification Analysis
- Regression Analysis
- Time Serious Analysis
- Prediction Analysis

Classification analysis

This type of data mining technique is generally used in fetching or retrieving important and relevant information about the data & metadata. It is also even used to categories the different types of data format into different classes. If you focus on this article until it ends, you may definitely find out that Classification and clustering are similar data mining types. As clustering also categorizes or classify the data segments into the different data records known as the classes. However, unlike clustering, the data analyst would have the knowledge of different classes or clusters.

Prediction Analysis

This technique is generally used to predict the relationship that exists between both the independent and dependent variables as well as the independent variables alone. It can also use to predict profit that can be achieved in future depending on the sale. Let us imagine that profit and sale are dependent and independent variables, respectively. Now, on the basis of what the past sales data says, we can make a profit prediction of the future using a regression curve.

Clustering Analysis

In Data Mining, this technique is used to create meaningful object clusters that contain the same characteristics. Usually, most people get confused with Classification, but they won't have any issues if they properly understand how both these techniques actually work.

Verify data availability:

My database is an located in open-source website of Kaggle, so I can download it and make some transformations whenever I want.

Define selection criteria:

As mentioned earlier, I use the database, based on the correctness and accuracy of the table. Also, if necessary, I can use the official ready-made ratings of the UEFA sports association, if necessary, to verify the correctness of the data.

2) Describing data:

The Transfermarkt market values are calculated taking into account various pricing models. A major factor is the Transfermarkt community, whose members discuss and evaluate player market values in detail. In general, the Transfermarkt market values are not to be equated with transfer fees. Numerous factors - the most important are listed below - calculate market demand for a player. Demand is defined using a paid transfer fee and salary in the context of the individual and situational parameters outlined below. It should be noted that in bigger leagues there is a heavier focus on transfer fees, while in smaller leagues where there is a greater emphasis on free transfers, the focus is primarily on salaries to determine market values. At the same time, market values are viewed both individually and in comparison, to other players/clubs/leagues. A player with an expiring contract has a transfer value of zero, the market value, however, is calculated independently from his actual transfer value.

In this respect, even a short contract length can only be taken into account to a limited extent. The individual transfer modalities are also relevant in the case of a possible difference between market values and transfer fees.

Most important factors:

- Future prospects
- Age
- Performance at the club and national team
- Level and status of the league, both in sporting and financial terms
- Reputation/prestige
- Development potential

- League-specific features
- Marketing value
- Number & reputation of interested clubs
- Performance potential
- Experience level
- Injury susceptibility
- Different financial conditions of clubs and leagues
- General development of transfer fees

3) Exploring data:

The database also contains the player's goals for the club and the number of matches played per season. When calculating the effectiveness and skill of a player by rating, it would be nice to take into account the realization of the player's dangerous chances in the match and the percentage of converted chances that turn into a goal. The name of the club and the competitiveness in the football club also depend on this, therefore it is necessary to take information from other public sources in order to clearly identify the strengths of the player.

4) Verifying data quality:

Most likely, I will use additional information on the Internet, in addition to the data in the database. Since for more accurate definitions, information only about goals, football clubs and matches will most likely not be enough.

Tasks

1) Create a model that will determine the highest quality player based on the data in the table.

2) Try to define an indicator that assumes high accuracy in predicting models across the entire class.

3) Run Estimator Standard Deviation measures the mean of squared errors, that is, the standard deviation between the estimated values and the actual value.

4) Extract all unnecessary data that does not take part in the implementation of the model

5) Show with an example the implementation of all the above methods

I'm planning to contribute with each task by myself cause I am only one person in my team.

Link to the repository: https://github.com/KirillRyrm/project2022