NBA Salary Prediction

Patrick Lomp
Artti Raasuke

Business understanding (Task 2)

Identifying our business goals

• Background

As we both are into sports and we both love to watch team sports. we wanted our project to be also sport related. Patrick has been following NBA for a long time and Artti is also a sports fan, who mostly follows football.

We in particular found NBA to be intriguing because it has been growing as a league and so are their salaries. This topic has been talked about a lot in media and how NBA has been growing a lot quicker compared to other major leagues like NFL. It would be interesting to know how much have salaries been growing compared to salaries in other fields and/or inflation and if we would be able to make a prediction model to accurately predict future salaries.

• Business goals

To give valuable insight to anyone interested in how much NBA salaries are growing year-to-year and what specific players might be earning in the future.

• Business success criteria

Make clearly understandable statements using the gathered data how the league's salary is growing.
Also we want our prediction model to be working at least somewhat realistically for salary prediction.

Assessing our situation

• Inventory of resources

The team – Patrick Lomp, Artti Raasuke. Both second year Bachelor's Informatics students.

Datasets - Two datasets. One dataset with players' statistics and another with players' salaries

Hardware – Both have access to personal PC's.

Software – We plan to use Jupyter, Tableau. Also different sklearn libraries for machine learning.

• Requirements, assumptions, and constraints

Schedule for completion – We have to be ready for poster presentation on December 19.

Acceptable finished work – Needs self-written code, Poster for the poster presentation.

• Risks and contingencies

Not having enough time – Our team needs to split work over the weeks given for working on the project, not do all the work on the last few day(s).

• Terminology

• Costs and benefits
Costs – 0 euros, no personal costs to us. We plan to use free software for our project.

Defining our data-mining goals
• Data-mining goals

Presentation – Need to make a poster for the final poster session.

Model – One of the goals is to make a prediction model, which predicts future salaries.

Report – Visualize our data and gathered information clearly and in an easy to understand way.

• Data-mining success criteria

Good data visualization – Assessment made by graders.

Model accuracy – We want the prediction model to be at least with the accuracy, which falls into -+15% range.

Data understanding (task 3)

Gathering data

In order to reach our goal of predicting the salaries of NBA players from the last century, we need performance stats of both players before and after 2000. For that we have a dataset called Seasons_stats_complete.csv, which has a lot of data about players and their stats throughout the years of their careers, starting from 1950. The older the data, the less information there is about the players. Another thing to consider is that the 3-point shot was introduced to NBA in 1979, so earlier statistics do not have that stat. Fortunately there is a statistic called player efficiency rating (PER), which takes into account all the positive and negative stats, and calculates an efficiency index. This index could be misleading in case of very small playtime, so such specimens might have to be removed from the dataset based on minutes played (MP). Fortunately, most of the rows of the players' statistics dataset have both of the required statistics, only 636 of 26063 rows don't have either PER or MP or either value, so the data loss is quite minimal.

In order to calculate the approximate salaries, we have a dataset which has salaries of NBA players from year 2000, called NBA_Full_Salaries_2000-2019.csv. This dataset has 37420 rows, out of which usable for us are 28074, since others don't have salary specified.

Describing data

The salaries dataset is very simple to use. It has 5 columns:

- row index,
- name of player,
- year of play,
- salary size and
- rank based on salary that year.

The performance stats dataset is a bit bigger with 50 columns, mostly containing various stats about player performance, but also about the represented team, player's age, field position and year of play. Most of the columns are named after the abbreviations of various statistics, so in order to understand them, we will be using various internet resources to understand them, such as https://stats.nba.com/help/glossary/ and https://en.wikipedia.org/wiki/Basketball_statistics. The main columns we will be focusing on are the following:

- year,
- player,
- MP (minutes played),
- PER (player efficiency rating).

We also might use other general statistics, such as total points thrown (PTS) or number of games played (G). One more thing to consider when joining tables based on names is that some names in the performance dataset have an asterisk (*) at the end of their name. This indicates that the player is in the NBA Hall of Fame.

Exploring data

Upon first inspections of joined data, some possible problems arise. First of such is that some players have represented multiple teams on the same year. This raises the question of which team paid the salary of the player. One possible solution would be to choose the most resultative of the bunch based on performance, or choose the one which has the most game time. Other than that, joining the datasets was very simple and gave us 10221 rows, but after dropping the duplicates we are left with 1673, which is a good starting point. The duplicates were dropped randomly, so this should be changed in the future.

Verifying data quality

We have a good sum of data, but the major problem is that some players have represented multiple teams on the same year, and since the stats are based on the team represented, it's difficult to choose from the lot of them. One possible solution would be to choose the most played team, which should be logical, but we will think through other solutions as well.

Project planning (Task 4)

Tasks

- Set up Github repository for maintaining and updating our project.
- Gather data and make clear goals what we are trying to achieve and how.
- Present our initial idea in practice and get feedback.
- Play around with our data to find useful knowledge.
- Plot NBA salary against inflation and see if there is a trend.
- Use machine learning to predict future NBA seasons salaries. Predict both average NBA player salary and salaries for specific players.
- Put it all together and visualize gathered information.
- Print our poster and present our topic at the poster session.

Methods and Tools

- Use Github to hold our codebase and data.
- Use Git and Github for version-control and tracking changes in development.
- Use Jupyter for writing code.
- Use Tableau to visualize gathered information.