U.PORTO

FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Machine Learning
Data Mining Project

Master in Informatics Engineering and Computation at FEUP, U.Porto

G53
André Costa (up201905916) - IF: 1
Diogo Fonte (up202004175) - IF: 1
Fábio Sá (up202007658) - IF: 1
Joaquim Monteiro (up201905257) - IF: 1

Domain Description

This project is a data mining case study regarding basketball tournaments. Our objective is to predict which teams are going to qualify for next year's playoffs, analyzing the past 10 years' data. The given dataset has data from players, teams, coaches, games and several other metrics.

Exploratory Data Analysis

Our first step was to analyze each table, to understand the provided data and reason about the features we could build with it.

We explored the meaning of each column, and identified which columns are useful and which aren't, and the relations between tables. This information was compiled in a Markdown document (`data_understanding.md`).

We dropped:

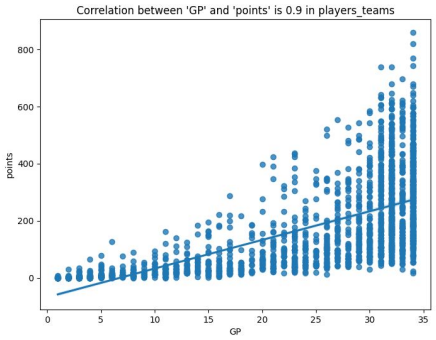
- columns with null values
- columns with always the same value (e.g. lgID)
- **columns that we found irrelevant after some analysis**

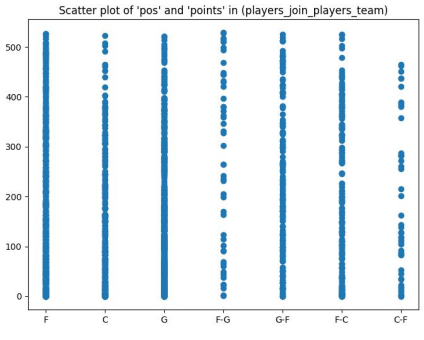
Exploratory Data Analysis

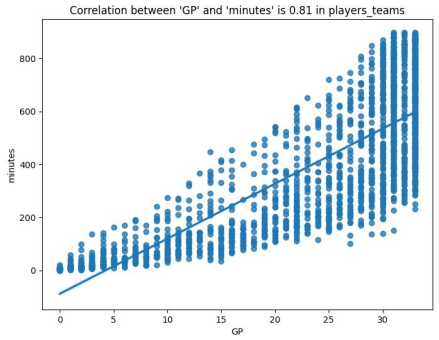
We made some plots to analyze the correlations between features that we characterize as irrelevant in each dataset and between features in joined data.

Examples of the most important analysis we made:

- Number of games played (GP) with player stats
- Position of each player related to its stats
- Irrelevant columns with same information







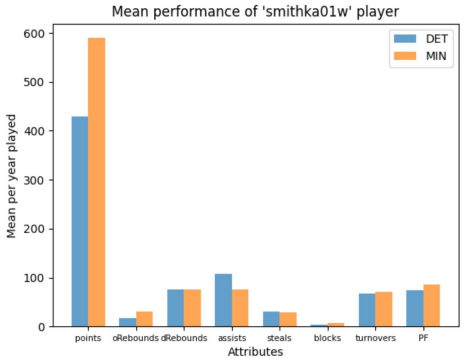
Data Mining Problem

How did we approach the data mining problem:

- Feature Selection
- Feature Engineering
  - Using some attributes to create better and more explicit ones
  - Aggregation of others attributes
  - Joining Tables
- Datasets creation
- Model Evaluation

Data Preparation

- Select the important attributes of each dataset.
- Combined the data between:
  - Awards_players and Players_teams
  - Awards_players and Coaches
  - Teams_post with Teams
- Series_post.csv and Players.csv were not used
- In the Player_teams we kept each player information (statistics of each season)
- In the Teams we added information about:
  - the season wins and losses, overall matches and win_rate and post_season too
  - total of championships won until the year of the row
  - total of awards of the players of the team on that year
  - the team offensive and defensive statistics
- In the Coaches we created new attributes:
  - coach_matches
  - coach_win_rate
  - coach_total_nrAwards



Data Preparation

More Feature engineering was performed on each of the combined datasets.

- Attributes were created:
  - Accuracy -> based on goals_made / goals_attempted
  - Overall win_rate until year X
  - nrAwards instead of name and type of award
  - number of championships based on the teams_post and ‘finals’ attribute of teams.csv
- Attributes were changed to its mean value:
  - points / GP -> like in the NBA official stats page
- Player_teams attributes importance and the mean approach of each attribute and its relevance to the problem.

Experimental Setup

For this phase, we started by creating each year dataset (excluding the first year).

Each dataset contains:

- Information of the pre-season
- Information created and selected during the data preparation phase of the previous year/last season

Then we selected the models we wanted to use, RFC and SVM.

Lastly, we make sure that there were exactly 8 teams, 4 of each conference, advanced to the playoffs in our results, by using the model's result probabilities as a ranking.

Data Mining Models

The models used were Random Forest Classifier and the Support Vector Machine.

For both of the models, we started by mapping the playoff value into boolean values, 0 as 'N' and 1 as 'Y'. Then, we used the model to perform the predictions. Following that, we used the following approach:

- As we are predicting the playoffs and not the ranking, we needed a way to only select 8 teams to the playoffs phase, 4 of each conference (East and West).
- So we sorted the teams by their probabilities of the model, and then we chose the 4 teams with the highest probabilities of each conference.

RFC:

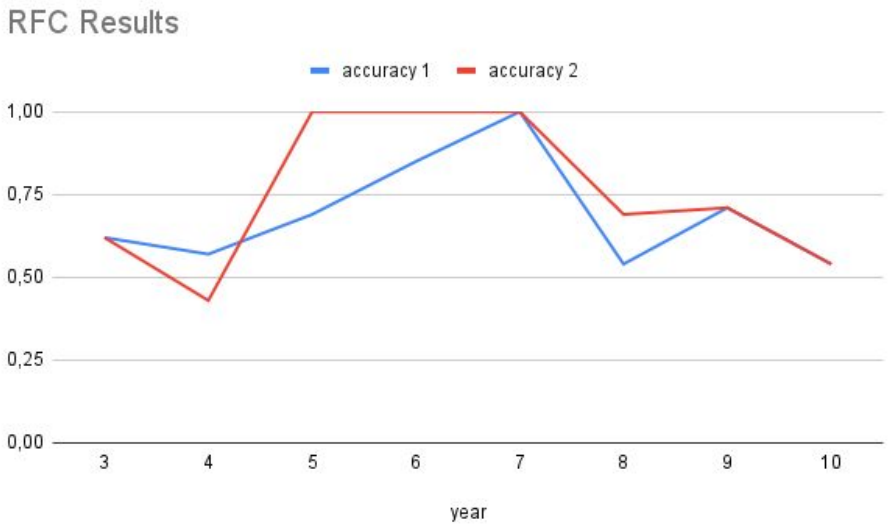
- This model had a good accuracy overall. We can see improvement over the years with both functions. However, for the last 2/3 seasons, the accuracy went down due to the fact that new teams appeared, and the model had less information to evaluate those.

SVM:

- This model had a bad performance, mainly when using only the last year as training data. It improved significantly with all previous years as training data, but still worse than the RFC model.

Results - RFC

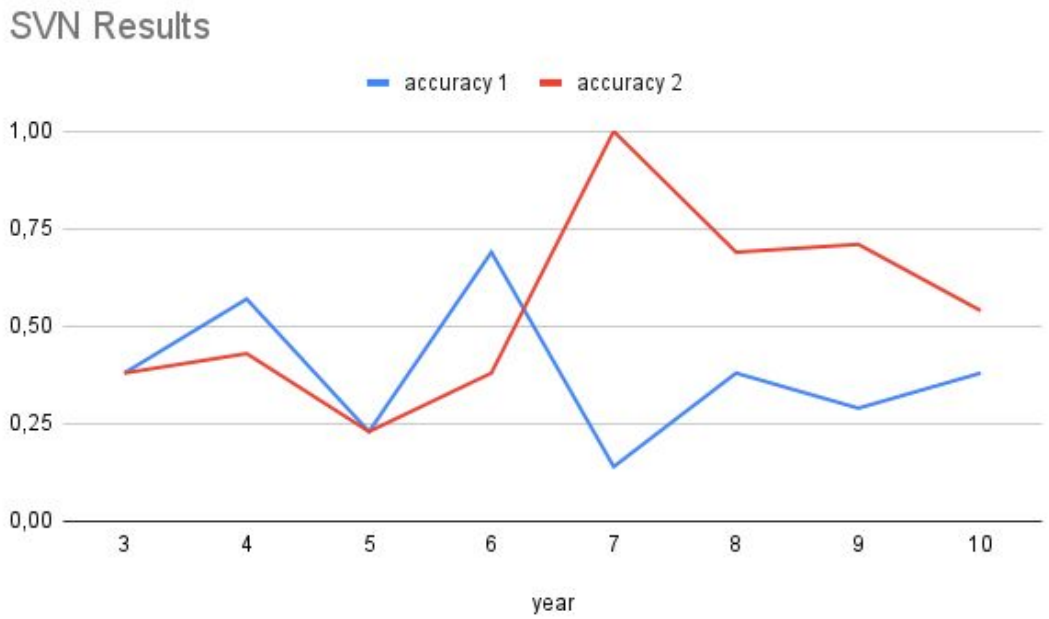
For each model, we created 2 different functions. One that only predicts with the previous year as training data, and the other function using all previous years as training data.



Accuracy 1 - Considering only the last year

Accuracy 2 - Considering all previous years

Results - SVN



Accuracy 1 - Considering only the last year

Accuracy 2 - Considering all previous years

Conclusions

Limitations:

- Important information missing (e.g. injuries that harm the performance of the players in the respective season)
- Problem with new teams and lack of information that affects the model predictions

Conclusions:

- We have learned the vital aspects of data mining methodologies and their practical application in predictive tasks, being the Feature Engineering one of the most important interfering the predictions results of the models.

Future Work:

- Different Players performance approaches:
  - mean performance of all seasons until of the year of the dataset
    - problem with season where players injured themselves and only played few games