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.

Data Understanding

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 attribute, and identified which attributes are useful and which aren't, and the relations between tables. This information was compiled in a Markdown document ('data_understanding.md').

We dropped:

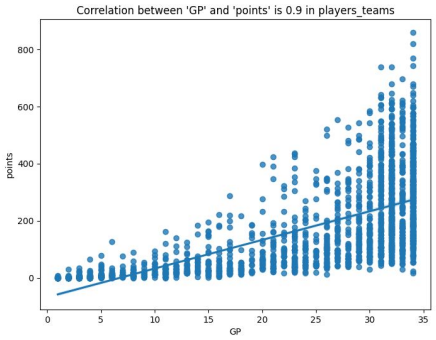
- columns with null values
- columns with always the same value (e.g. lgID)
- **columns that we found irrelevant through some analysis**

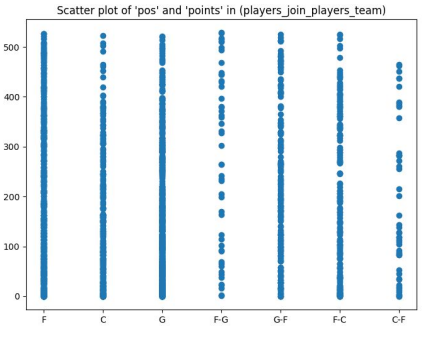
Exploratory Data Analysis

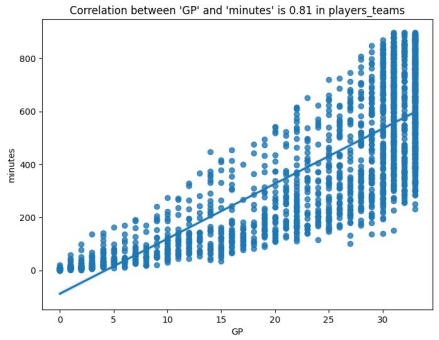
We generated plots to explore correlations among features, focusing on identifying redundant information, irrelevant features relatively to the player/teams statistics and the significance of attributes related to the evaluation metric.

Examples of our analyses:

- Number of games played (GP) with player stats
  - We can conclude that there is a pattern/correlation when the ‘GP’ increases the points increases too (also depending on the mins played by each player)
- Position of each player related to its stats
  - Most of player stats, don't have a high correlation with its position
- Columns that provide the same logic/information (e.g. ‘GP’ and ‘minutes’)

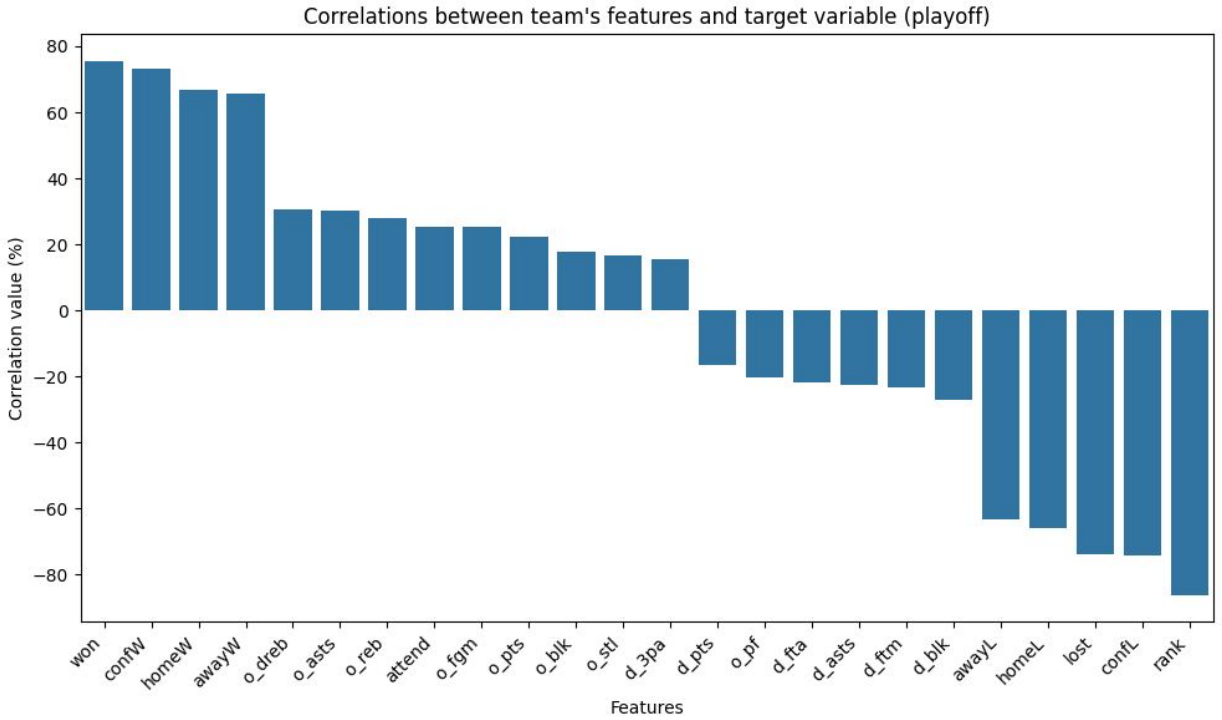






Exploratory Data Analysis

We also analyzed the correlation of known features with the Target variable, *playoff*. This allowed us to identify which features contribute most effectively to the model's decision and those that, due to lacking any relationship, could be discarded.



Data Mining Problem

We started by selecting the features that we do not found irrelevant through the ‘Data Analysis’ phase.

Followed by the Feature Engineering, we applied some of its techniques:

- Creation of new attributes based on other(s) ones.
- Aggregation of information of several attributes into new ones.
- Joining information of multiple Tables.

Datasets creation for each year with the information of the new year’s pre-season and the previous years.

Model selection and evaluation of its results.

Data Preparation

Combined the data between:

- Awards_players into the Players_teams.
- Awards_players into the Coaches.
- Teams_post with Teams.
- Series_post and Players were found to have irrelevant information for the prediction of the playoff.

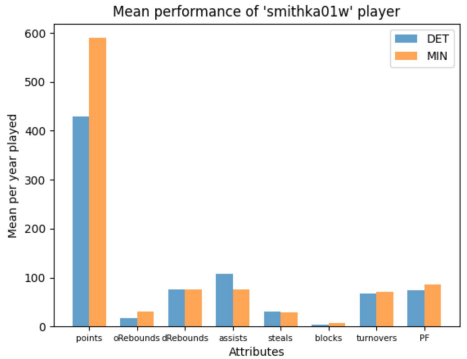
First approach: we focused on the Player_teams information, keeping each player statistics of each season. This approach allowed us to understand its importance, due to the fact of the rotation of the players of a team throughout all the seasons.

- It's important to notice that with this approach we avoided cases where the investment of a team increases/decreases significantly one year and its rank also varies significantly. Without the investment's information, this approach would give better results.

Data Preparation

Second Approach: Through more analysis, we noticed nuanced variations in player performance across different Teams. Despite the absence of important information of a Team, we conducted a meticulous analysis by aggregating data from each season, focusing on individual team statistics.

The processed information encapsulates the average performance metrics of new players for each team, along with the mean statistics per game for every team. Additionally, we integrated insights into the coaching performance over the years.



Experimental Setup

Our **Final Dataset** for each season comprises both pre-season and last season information. This comprehensive dataset also integrates attributes calculated based on historical data from all the years leading up to the current year, such as the win-rate of the team and coach.

Then, we proceeded to the Model Evaluation phase. Employing the **Random Forest Classifier** and the **Support Vector Machine** models, we sorted the teams based on their respective model probabilities. Subsequently, we strategically selected the top four teams with the highest probabilities from each conference, obtaining the 8 teams of the playoffs phase.

Data Mining Models

In developing our models, we implemented a dual-function approach for each model. The first function exclusively utilizes the previous year's data for training, while the second function incorporates information from all previous years. This extended approach resembles a variation of the **sliding window technique**, demonstrating superior performance compared to models trained with only a single year of data.

Notably, the Random Forest Classifier (RFC) outperformed the Support Vector Machine (SVM), particularly when employing only one year of training data. The SVM results in this scenario were notably suboptimal.

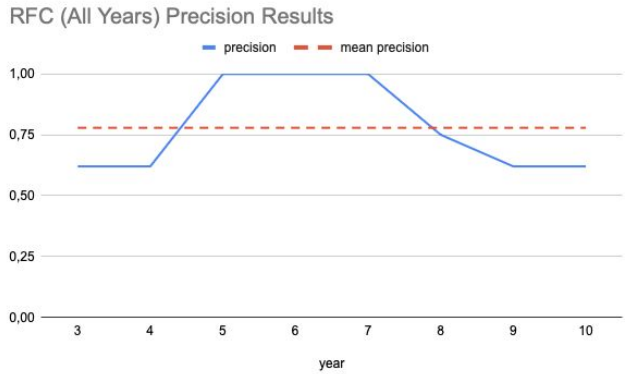
For both models, we assessed performance through key metrics, including the confusion matrix to identify areas of weakness, feature importance analysis, ROC and Learning Curves for visual insights, and cross-validation scores to gauge overall model precision.

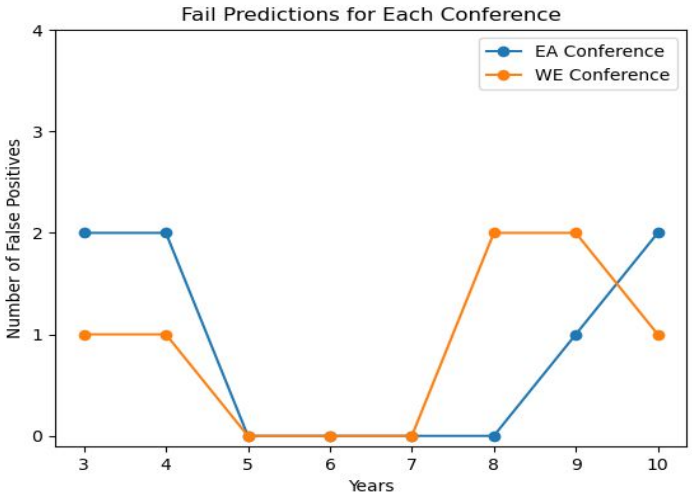
Results - RFC

Analyzing the approach that yielded the model with the highest **Mean Precision of X**, we observe years exhibiting 100% precision and suboptimal results in specific instances, notably during years 3, 4, 9, and 10.

The challenges encountered in these years stem from missing information, such as the introduction of new teams, substantial investments in some teams, and the absence of data concerning players in their first year.

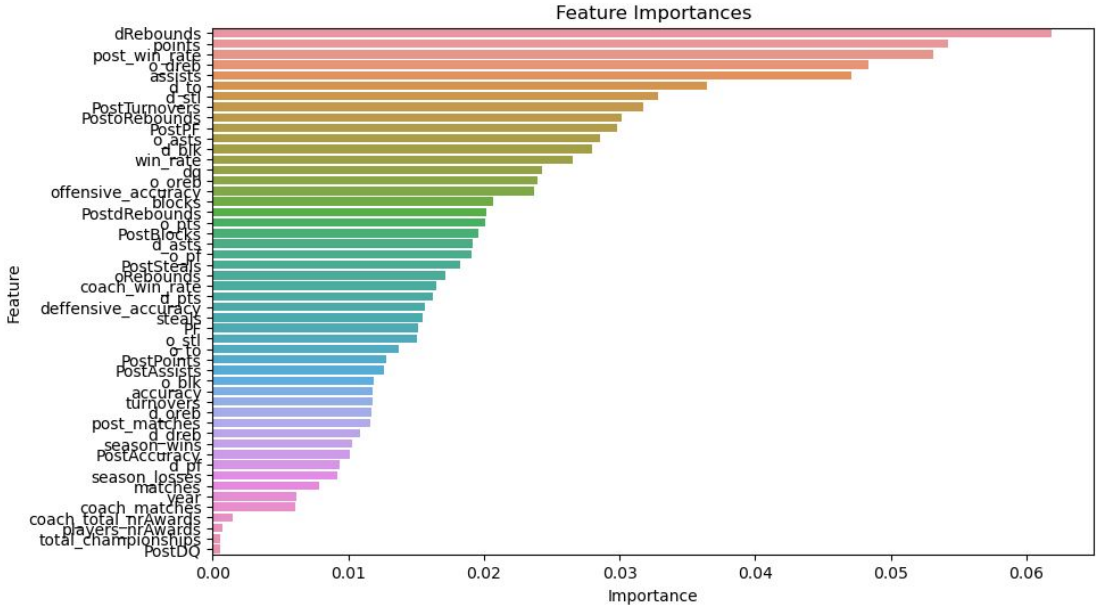
Using the **confusion matrix** to determine how many **false positives** the model had per year, we can analyze how many fail predictions it had for each conference, providing significant information for the client viewpoint of each conference.





Results - RFC

The analysis of Feature Importances reveals opportunities for enhancements in the feature engineering phase, indicating viable paths for increasing overall model performance.

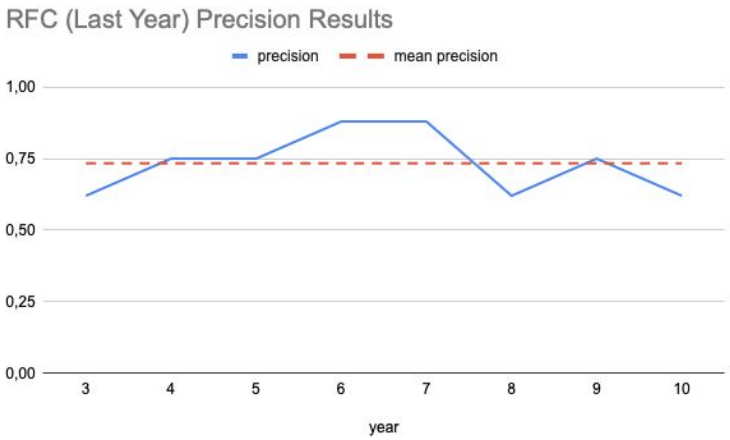


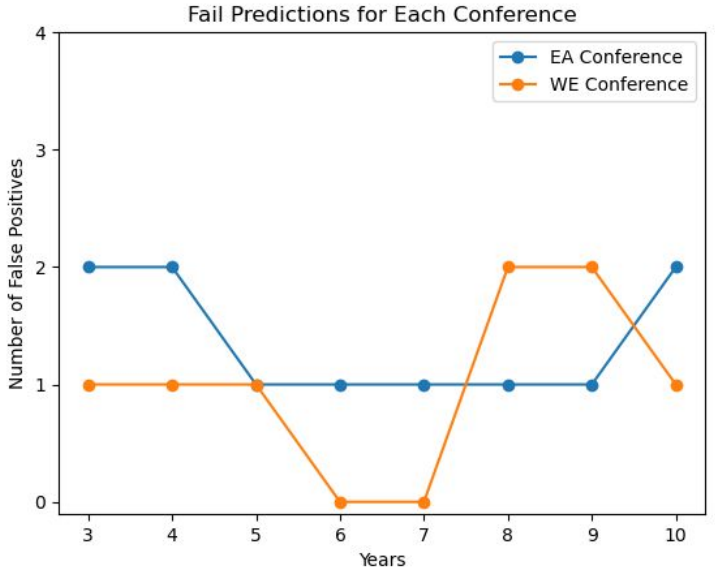
Conclusions

We have learned the vital aspects of data mining methodologies and their practical application in predictive tasks. Notably, we've recognized the importance of Feature Engineering, interfering the predictions results of the models. Understanding the impact of feature manipulation on model predictions is integral to achieving accurate and reliable outcomes.

Our findings highlight the applicability of these approaches in the context of real-world applications, particularly in the betting business. The Mean Precision of X, highlights the potential for using data mining techniques in this sector.

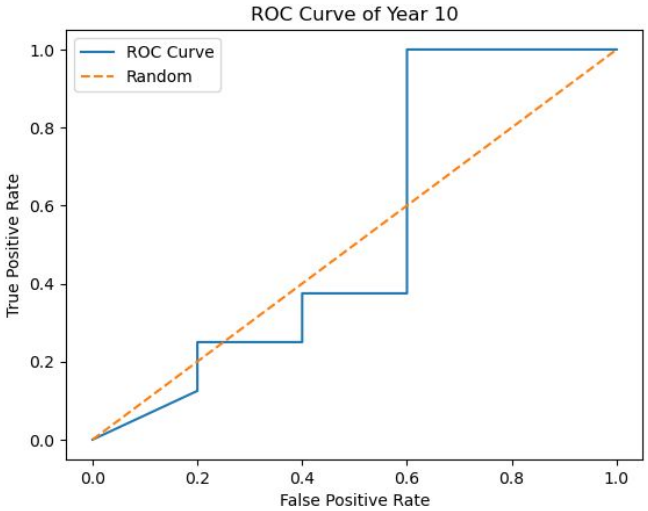
Annexes

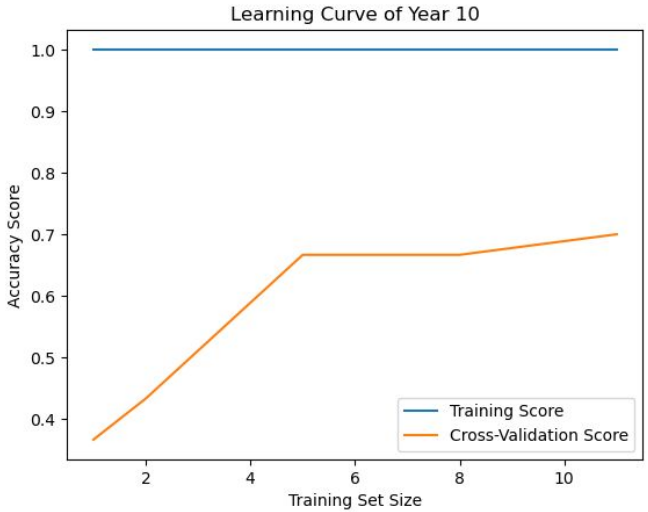

RFC (Last Year) Precision Results
precision
mean precision
1,00
0,75
0,50
0,25
0,00
3
4
5
6
7
8
9
10
year


Fail Predictions for Each Conference
EA Conference
WE Conference
Number of False Positives
4
3
2
1
0
3
4
5
6
7
8
9
10
Years

Annexes

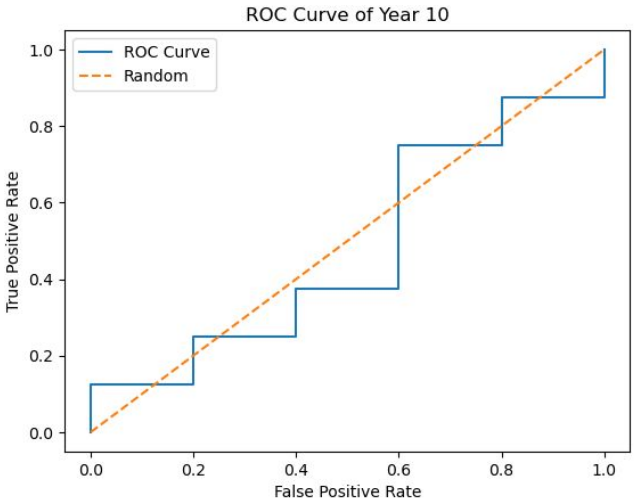
RFC (last year) ROC and Learning Curves

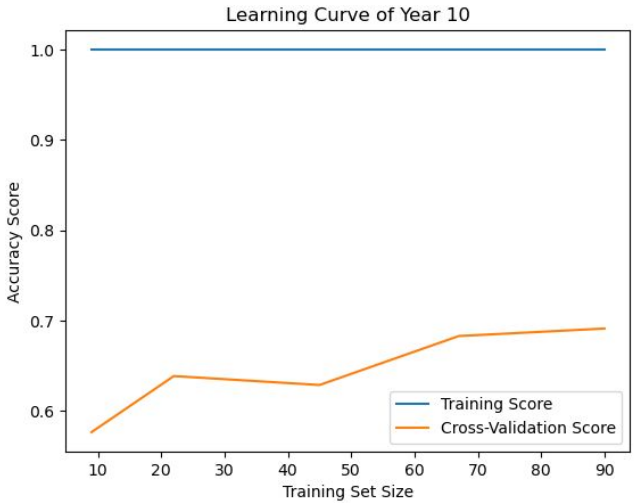




Annexes

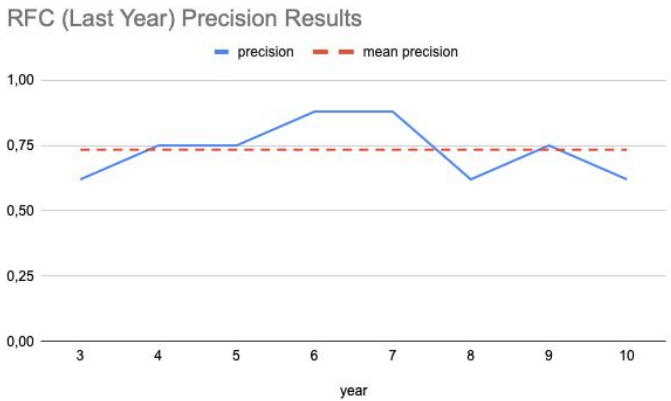
RFC (all years) ROC and Learning Curves

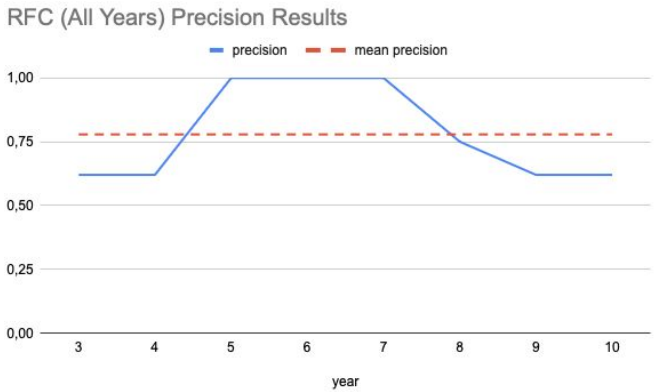




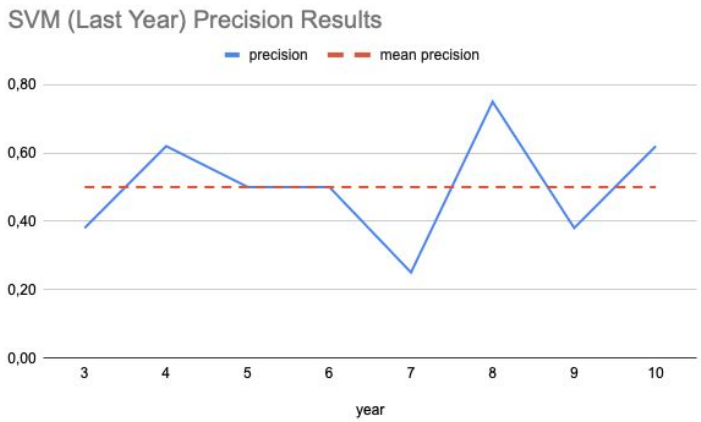
Annexes

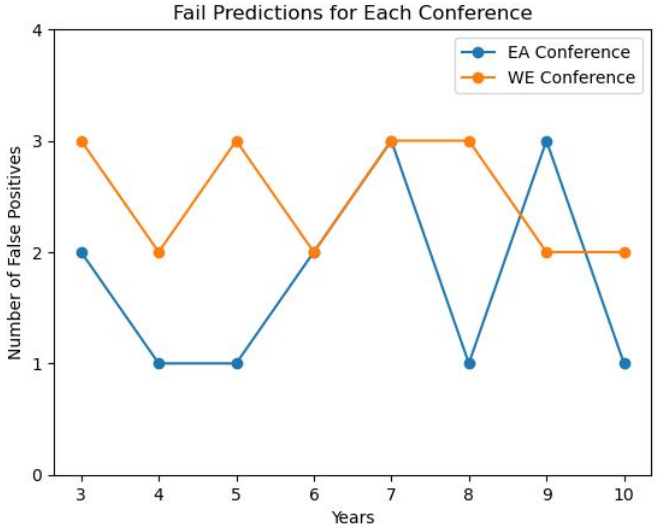
RFC last years vs all years precision





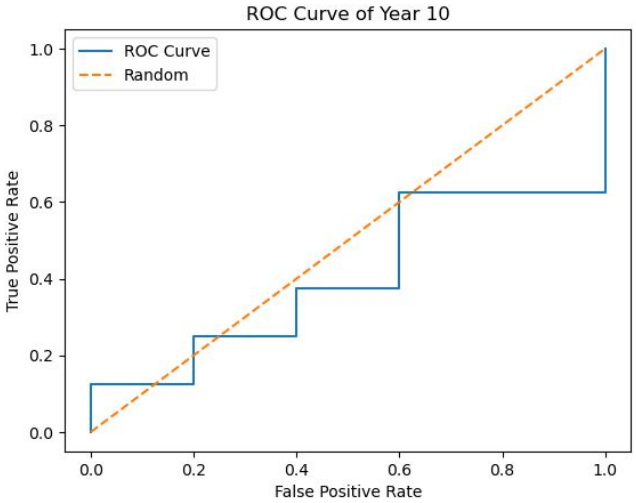
Annexes

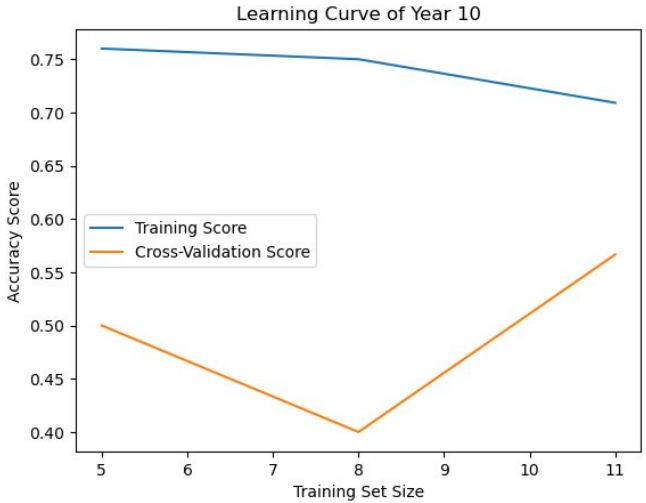

SVM (Last Year) Precision Results
precision
mean precision
0,80
0,60
0,40
0,20
0,00
3
4
5
6
7
8
9
10
year


Fail Predictions for Each Conference
EA Conference
WE Conference
Number of False Positives
4
3
2
1
0
3
4
5
6
7
8
9
10
Years

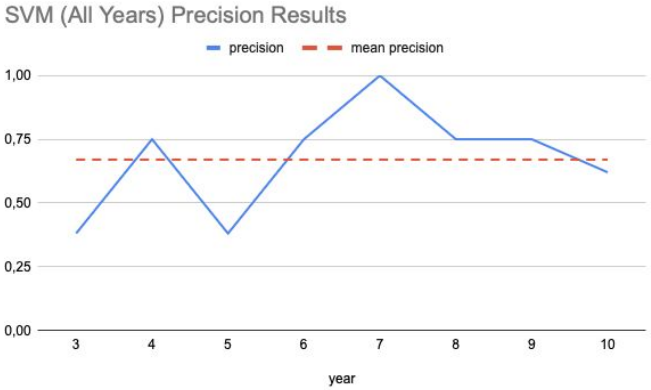
Annexes

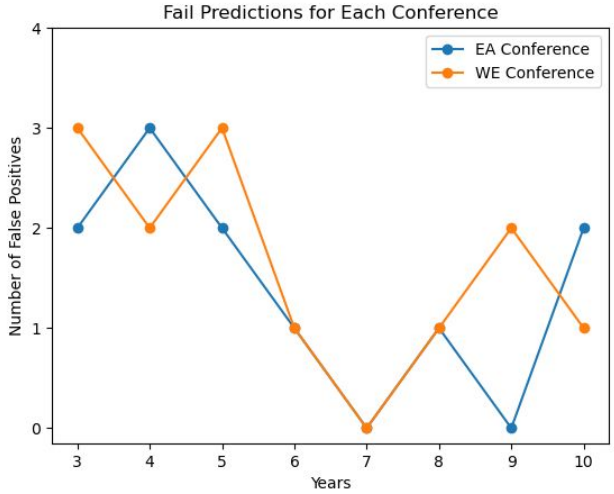
SVM (last year) ROC and Learning Curves





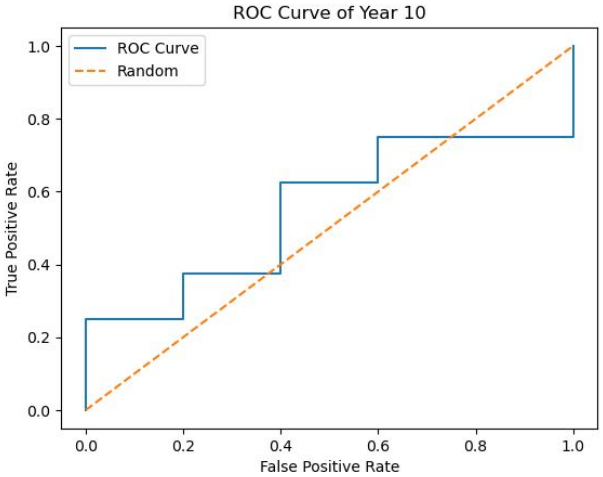
Annexes

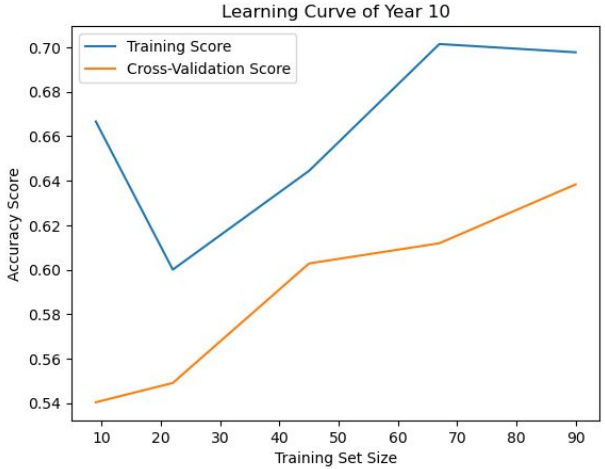

SVM (All Years) Precision Results
precision
mean precision
1,00
0,75
0,50
0,25
0,00
3
4
5
6
7
8
9
10
year


Fail Predictions for Each Conference
EA Conference
WE Conference
Number of False Positives
4
3
2
1
0
3
4
5
6
7
8
9
10
Years

Annexes

SVM (all years) ROC and Learning Curves





Annexes

SVM last year vs all years precision