Knowledge discovery - NBA 2021/2022 dataset

Nina Masaryková, Marek Štrba

March 2022

1 Dataset

The dataset we chose to use for our project can be found on the following link https://www.kaggle.com/vivovinco/nba-player-stats the dataset is being updated as the games are played. We have downloaded our dataset on 17.02.2022, right before the all-star weekend which took place from 18.02.2022 until 20.02.2022, therefore this version of the dataset was perfect for our hypothesis described further down the line.

Data structure

The dataset itself contained 734 records. Every record represents personal statistics of one player. Some of the statistics are per game, other are counted as totals, more on that in the next section.

Some of the players were traded to another team during the season. These players have multiple records. One where the team is noted as TOT (total) and one for every team they played in. The total record aggregates the season statistics while the team records mark only the statistics achieved while playing for the given team. Every record consisted of 30 columns - 26 numerical and 4 categorical. We added our own column - ALLSTAR to mark whether the player was selected for the all-star game or not. We added it manually according to the official NBA website (https://www.nba.com/allstar/2022/all-star-roster). The value "1" marks an all-star, while the value "0" marks the rest of the players.

The data gathered in this dataset were collected over the span of 60 games period (the maximum number of played games by a player was 60).

Column cheat-sheet

Basketball statistics can be unclear for someone who is not familiar with the sport, we therefore prepared a cheat-sheet (Table **??**) to explain the columns of the dataset we used. We omitted explain the obvious columns such as name or team.

2 Data preparation

The dataset which we picked was already in great condition. The nature of the data ruled out any measurement errors, or errors caused by a faulty machines. We decided to make only few augmentations. We decided to remove the records which contained the statistics of traded specific by the team they played in. Thankfully we didn't have to calculate the aggregated statistics, because they were already provided as the part of the dataset. We decided to drop the partial stats, because we didn't aim to do any predictions based on the change in performance or the trades themselves. The other significant step we made was removing all records of players who played less than 10 games. We considered these records to be not objective enough since the game pool of less than 10

games could significantly deviate from the overall performance of the player and could show a good player in bad light or a bad player in a good light. Also comparing statistical averages between themselves when one was achieved over the course of 5 games and the other one over the course of 50 just isn't relevant. Other than the already mentioned adjustments, we only fixed some of the names of the players which were broken by the UTF encoding.

3 EDA

After we removed the duplicates from the dataset caused by the traded players records, we have checked the dataset for duplicates and proved that there were none left. Next we checked the dataset for any missing values. We found no missing values thanks to the aforementioned nature of the dataset. Our first statistical analysis over the data was creating a heatmap of correlations between all possible pairs of the attributes (Figure 1). In this step of the analysis, we knew we wanted to predict scored points per game and whether a player gets selected as an all-star or not, therefore we mainly focused on the correlations of the PTS and ALLSTAR attributes.

Regarding the points prediction, it would be an easy process to pick some of the statistics based around the Field goals or 2P and 3P attempts and predict the number of points based on them. The FG and FGA both even have a correlation of 1 with the PTS attribute, we however deemed it too straightforward and not really a problem to be solved using an ML model. We therefore delved deeper in the other attributes and found interesting correlations with PTS attribute, which weren't the first to come to mind. We chose the following attributes for further analysis: MP (0.9 corr), FT (0.9 corr), AST (0.7 corr), STL (0.6 corr), TOV (0.8 corr) and PF (0.6 corr). All these attributes make sense when connected to the PTS prediction because they all describe the player as a player who gets enough time on the playing court (MP) and during this time they often have the possession of the ball (AST, TOV), are active when the opposing team has the possession (STL, PF) and are often a scoring danger for the opposing team which needs to be stopped by a foul (FT). A player which would fit this description is surely very likely to score high amount of points every game.

Choosing the attributes for the ALLSTAR attribute prediction proved trickier. The highest correlation we were able to find was 0.6 of FT. We decided to not use this attribute, because we deemed attribute PTS which covers the point production as a whole more descriptive of the overall level of the players' performance. The PTS attribute itself had the correlation of 0.5 which we deemed as not as big of a drop, in exchange for the nature of the attribute. We also discussed whether to use the GS or G attribute. Once again GS attribute has a slightly higher correlation (0.3 compared to 0.2), however the fact that usually not more than 2 starting players from a team are selected for an all-star games left us with the fact that we would "boost" the viability of the 3 of the remaining starters. Other than that the G (games played) is one of the key decision points for the voters when they pick the all-star players. To secure

that the G stat which can be high even for a player not suitable for being an all-start will not drastically deviate the results from truth, we added the MP attribute which shows how much time did the player spent in play in the games they played and should therefore eliminated roleplayers (weaker players from the team) from being falsely marked as all-stars. The last attribute we wanted to analyse for this prediction was the eFG% which by it's definition should be one of the attributes which describe the player's level of play with a higher rate of accuracy.

After this initial analysis we focused on the specific attributes we marked as potentially viable predictors.

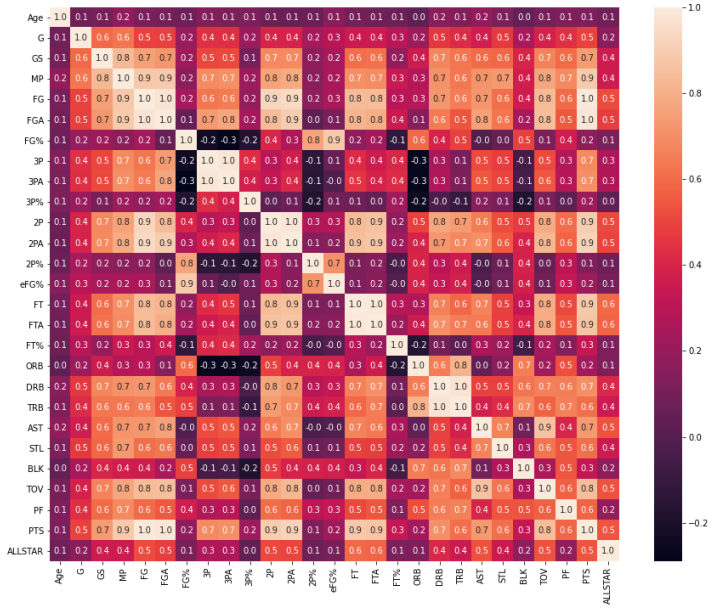


Figure 1: Heatmap denoting all the correlation ratios between any two given attributes found in the dataset.

Analysis of attributes in the scope of hypotheses

In this step we took the attributes one by one and used methods such as QQ-plots, Pearson's correlation, Shapiro-Wilk test and Boxplots to determine, whether the attributes could prove viable as predictors or not.

Column name	Correlation with PTS	p-value (Shapiro-Wilk)
PTS	1	0.00000000000000014610802708830
TOV	0.847	0.00000000000000000001182280590
STL	0.611	0.00000000003756198824200075157
PF	0.581	0.01528532616794109344482421 88
MP	0.885	0.00000007207547447762863 1575
FT	0.881	0.00000000000000000000000052660
AST	0.737	0.00000000000000000000002948490

Table 2: Statistical analysis of attributes related to the points prediction

Attributes for the points prediction

As the first step we analysed the PTS attribute itself and arrived to a conclusion that based on the QQ plot shown on Figure 2 and the p value calculated by the shapiro-wilk test (Table) the PTS attribute is not normally distributed.

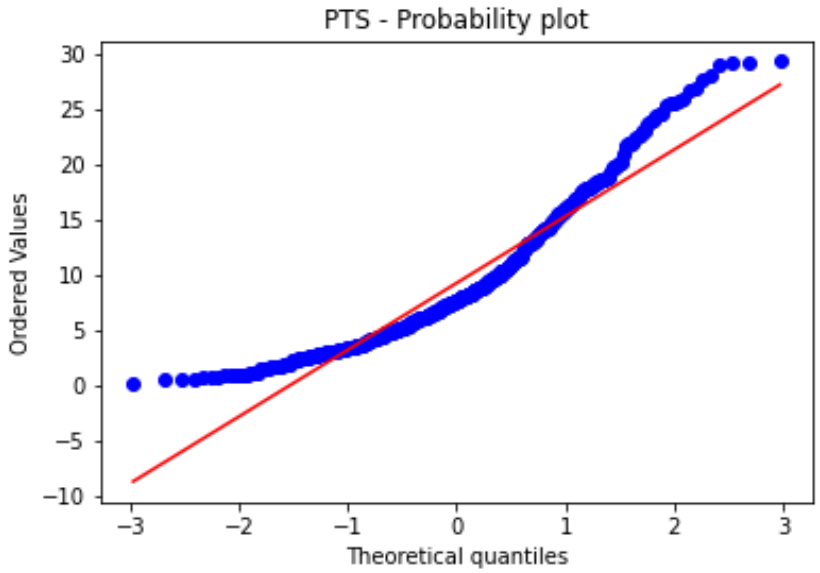
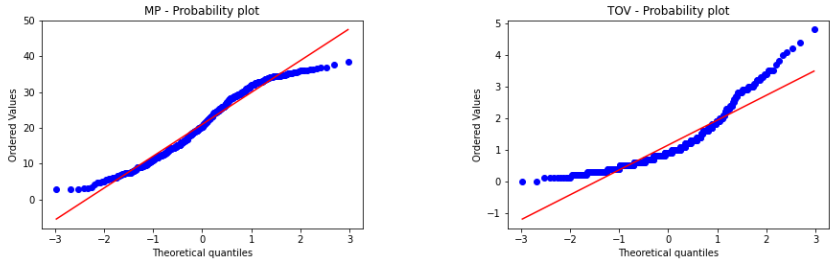


Figure 2: QQ plot of the PTS attribute

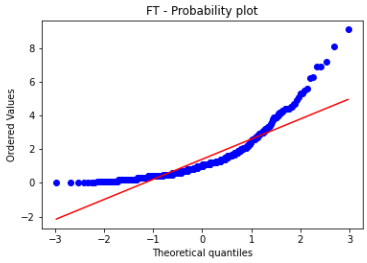
Similarly to the PTS attribute, we plotted the rest of the attributes in a QQ-plot as well (Figure 3). For the sake of clarity we included only the plots of attributes we decided to use in this report, all plots are available in the attached jupyter notebook.

We decided to drop the STL and PF attributes from the predictor pool mainly because their distribution in relation to PTS shown on the scatter plots

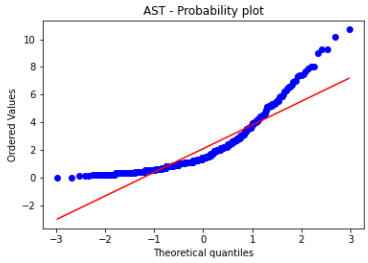


(a) MP

(b) TOV



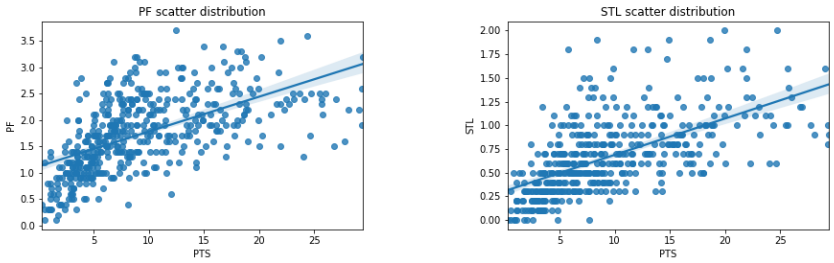
(c) FT



(d) AST

Figure 3: The QQ plots of the selected attributes for the point prediction. These QQ plots visually confirm the results of the Shapiro-Wilk test and show that none of these attributes is of normal distribution. We tried transforming them using the Yeo-Johnson transformation, however it did not help in a significant way.

bellow (Figure 4)could only hardly be approximated using a regression model without a significant error or a very high risk of over-fitting.



(a) PF (b) STL

Figure 4: The scatter-plots denoting the distribution of PF (a) and STL (b) attributes in relation to PTS.

The relations and distributions of the selected predictors as well as the PTS attribute can be seen plotted on the scatter matrix bellow (Figure 5)

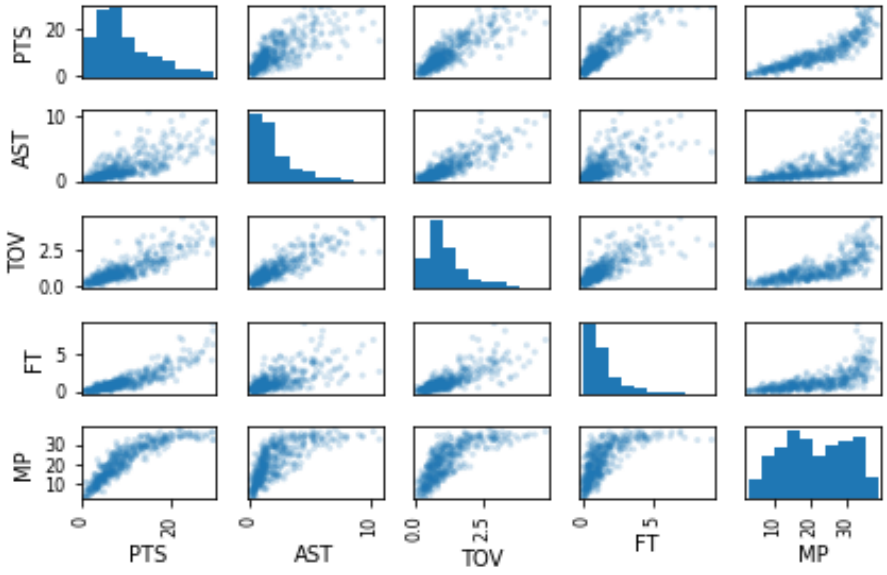
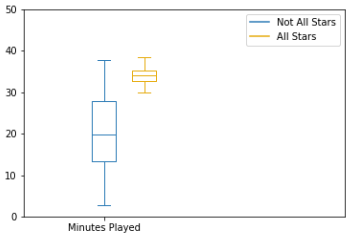


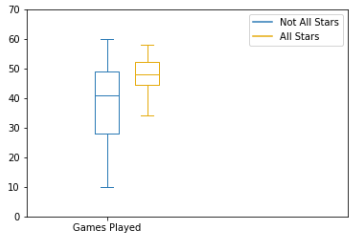
Figure 5: Scatter matrix describing the relations between selected predictors and the PTS attribute

Attributes for the all-star prediction

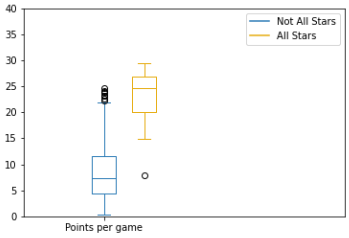
The attribute selection for this prediction was rather different than in the case of the points prediction. Our aim was to select the attributes which could be used to split the all-stars from the rest of the players with the highest possible accuracy. To determine this we used Boxplots (Figure 6) and Histograms (Figure 7)where we differentiated between the all-star records and the other players.



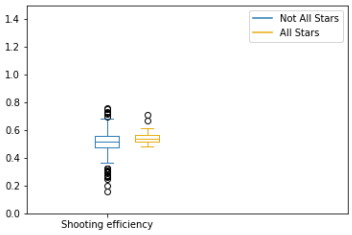
(a) MP



(b) G



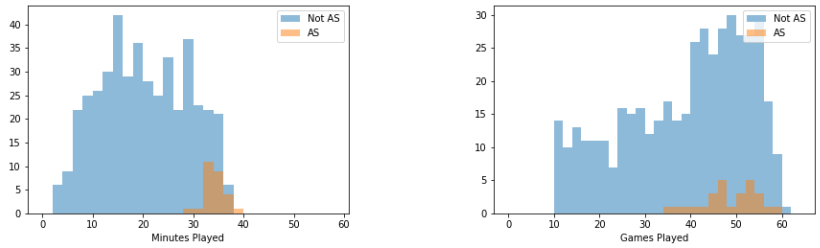
(c) PTS



(d) EFG%

Figure 6: These Boxplots show the splitting potential of the given attributes between all-stars and other players. Minutes played (a) show a clear distinction between all-stars and other players. Similarly the Games played (b) show the difference, however this time it is not as clear as with the minutes. Points per game (c) also show a clear difference, however the outliers of the non-all star players overlay with the all-star level production. We do not consider this to be a problem, since it only means there are players, which deserved to be selected to the all-star game, but were passed on due to the limited spots on the roster. The final attribute eFG% failed to provide a clear cut between the all-stars and other players.

Based on the results of the analysis using the Boxplots and Histograms, we decided to move forward without the eFG% attribute, which despite having a great promise based on it's nature failed to provide a clear distinction between the all-star players and the rest of the player base.



(a) MP

(b) G

(c) PTS

(d) EFG%

Figure 7: The displayed histograms only confirmed the findings made by interpreting the Boxplots above.

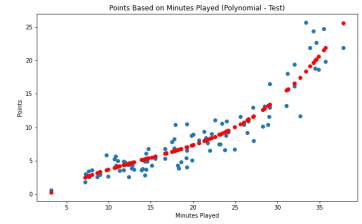
Creating the hypotheses

Our aim from the beginning was to have one hypothesis for a regression problem and one for a classification problem, which we would test using two different models. Even from the early stages of the analysis we knew we would aim to predict the points per game stat using the regression and the all-star selection using two different classifiers. After the analysis of the attributes we arrived to these two hypotheses:

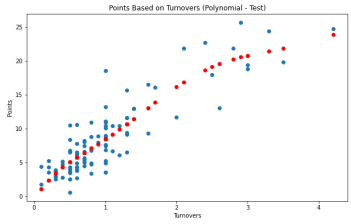
- Predicting PTS based on the MP, TOV, AST and FT statistics
- Classifying the players to all-stars and regular players based on PTS, MP and G statistics

4 Points prediction

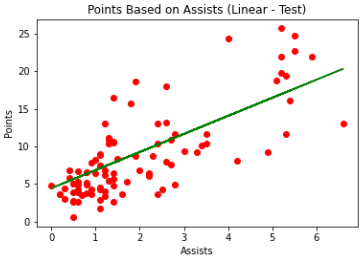
With all predictions we split the dataset to training and testing (validation) sub-datasets in the ratio of 80:20. Our first step was to analyse the predictors separately and try to fit the best regression based on their distribution. The best models for each attribute can be seen bellow (Figure 8) to see the results of the model performances refer to Table 3.



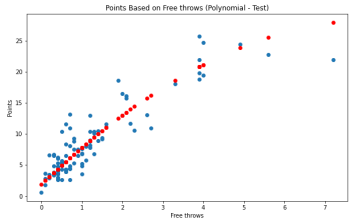
(a) MP



(b) TOV



(c) AST



(d) FT

Figure 8: The plots above display the models with best results for every predictor by themselves

After analysing the results of the partial predictors, we lost some of the confidence we had for the AST attribute as a predictor, as seen both on the plot and the results in the table, it struggled with predicting the correct point values for the players which had higher number of assists.

We moved to the complex multi-attribute model. First we tried the linear regression with all attributes which yielded following coefficients: -0.03616471 (AST), 0.92043426 (TOV), 2.1798644 (FT), 0.33266538 (MP). These results lower our the value of AST even further. Interesting fact however was that adding 1 made free throw to players statistics, which is worth only 1 point should raise the total points by more than 2.

We tested multi-attribute models using the same steps as with the single attribute models. As the first step we tried the linear model and then we tried the polynomial models of degree 2-7 and recorded the best polynomial result with the used degree. The multi-attribute groups we used were (i) all pre-selected potential predictors (TOV, AST, FT and MP) and (ii) all predictors except for the under-performing AST.

All of the results are recorded in the Table 3. The model with the best performance proved to be the **Polynomial regression of 2nd degree with the usage of attributes MP, TOV and FT as predictors**. This model had the root mean squared error (the average difference between real PTS and predicted PTS) of only 1.560 which is when we take in the consideration the average PTS of 9.355 rather satisfactory.

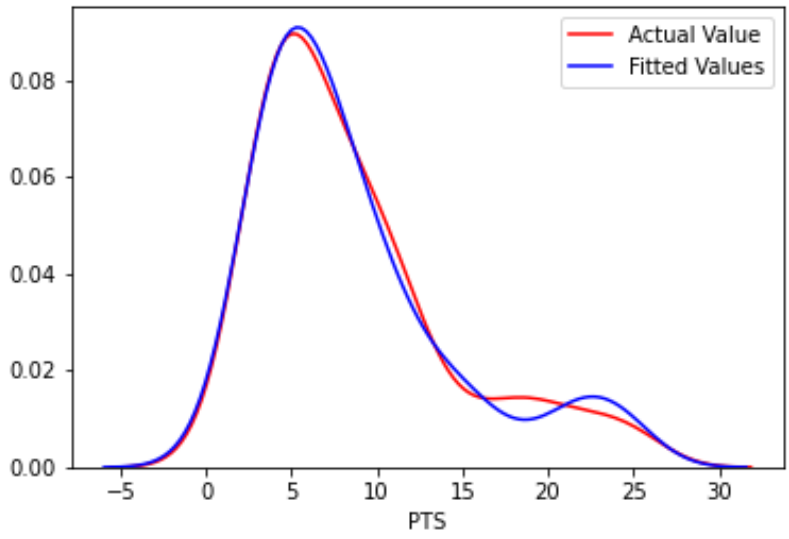


Figure 9: This plot shows the comparison of all PTS values predicted by the best model and all real PTS values

We explain the failure of the AST as a predictor with its rather wide cone distribution in the relation to points, which could be seen in Figure 5. The AST proved a valuable predictor for the players which had low score in this statistic, however the players with a high number of assists per game proved to be difficult to predict. We explain this mainly by the fact that the players who score many assists per game are most likely either pure facilitators (players focusing on assists) or they also score a high number of points (all-around offensive players). These two groups will be on the opposing sides of the spectrum when taking the PTS into consideration, however they will be very close in the AST numbers, therefore creating the hard to fit cone.

If we would want to lessen the error of the prediction we would add predictors such as FG, 2P or 3P which are tightly bound to the PTS total itself, however that seemed boring and not fun.

Used attributes	**Lin MSE**	**Lin RMSE**	**Best poly degree**	**Poly MSE**	**Poly RMSE**
MP	6.633	2.575	3	3.969	1.992
AST	14.278	3.779	3	14.662	3.829
TOV	7.819	2.796	6	7.458	2.730
FT	7.136	2.671	2	5.702	2.388
ALL	2.506	1.583	2	2.751	1.659
No AST	2.539	1.593	2	2.434	1.560

Table 3: The results of every tested attribute set on a linear model as well as on the best performing polynomial model. The best performing was the polynomial model of degree 2 with the usage of all predictors, except the AST attribute. The worst performing predictor was the polynomial model with only the AST attribute.

5 All star prediction

Same as with the regression problem we divided the dataset to training and testing (validation) sub-datasets in the ratio of 80:20. Our testing sub-dataset contained 90 non-all-star players and 4 all-star players. We kept this disproportion of the classes in mind when we rated the success of the different models, therefore we focused on the accuracy metric and the area under the ROC curve when comparing the models between each other. It is also important to note, that the all-star selection process is rather subjective, because it is based on the popular vote, player's vote, the vote of journalists, head coaches and the commissioner of the league. This fact cannot be accounted for in any way using the data about the player's performance. Therefore it is a fact that some of the players voted as an all-star will be classified by the model as a non-all-star and the other way around. A miss-classification of some level is inevitable based on this fact.

We started with the logistic regression and analysed the single metric models first. The worst performing was the model based on the G (Games played) attribute. We explain this by the fact that the Games played as a stat aren't

directly connected to the quality of the players performance, and the G attribute as seen in the heatmap mentioned before (Figure 1) correlates with the ALLSTAR attribute only by 0.2. We however, aimed to use this attribute in a combination with minutes played to target the players which played the most minutes in the highest number of games, which shows their value to the team. To fully test whether the G attribute is detrimental for the combined predictor or not, we added a model which uses MP and PTS only. The best logistic regression model proved to be the model which used all chosen predictors. The model with the G attribute omitted scored the same on the AUC, however when we refer to the Table 4 we can see that it was less accurate on predicting the non-all stars when compared to the model which uses all the attributes.

After we created all logistic models we moved onto SVM models. SVM models are specific in their use of a so-called kernel function, which is used to transform data into a higher dimension, which allows SVM to classify the data more accurately using a hyperplane. We ran a cycle which tested the SVM on given attribute with RBF, Linear and Sigmoid kernels and we picked the one which performed the best by the accuracy. RBF proved to be the best one for all attribute combinations, therefore all SVM results mentioned here are from a RBF kernel function. We tested the same combinations of parameters as we did on the logistic regression. It was interesting that according to the AUC metric only the Games based SVM performed better than it's logistic counterpart. Minutes and Points both performed worse. The combined models however, both performed better than their logistic counterparts. The best model overall was **SVM classifier with MP, PTS and G attributes** which predicted the classification with a 98% accuracy and scored 0.9805 in the AUC metric.

All ROC and AUC metrics can be found in the plot bellow (Figure 10).

Attributes used	Model type	TP	TN	FP	FN
PTS	Logistic	2	89	1	2
G	Logistic	0	90	0	4
MP	Logistic	0	89	1	4
ALL	Logistic	2	90	0	2
PTS, MP	Logistic	2	88	2	2
PTS	SVM	2	89	1	2
G	SVM	0	90	0	4
MP	SVM	0	90	0	4
ALL	SVM	2	90	0	2
PTS, MP	SVM	2	89	1	2

Table 4: Results of all classification models. TP - true positive (all-star classified as an all-star), TN - true negative (Non-all-star classified as a non-all-star)

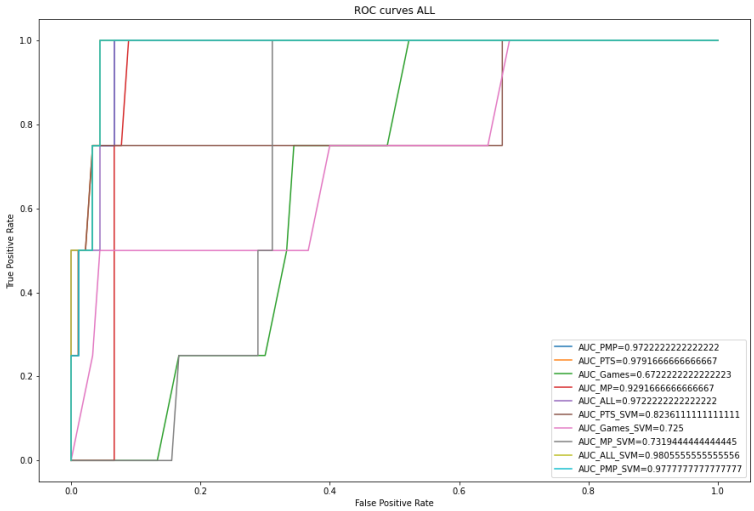


Figure 10: All ROCs and AUC scored based on the used models.