UEFA EURO 2016 Tournament Predictive Modeling

Atanas Mirchev Georgi Dikov

Statistical Modeling and Machine Learning
8th July 2016

Modeling Concept

- Random Forest as a baseline
- Multivariate Logistic Regression with categorical dependent variable
  - Win/Draw/Loss classification + confidence
  - Interpretation of the co-variate's importance
  - Simple to implement, successful
- LogRegression with hard “draw probability” cut-off
  - Predict draw or not-draw
  - If not-draw then predict loss/win
- Other models for comparison: SVM, Poisson Regression

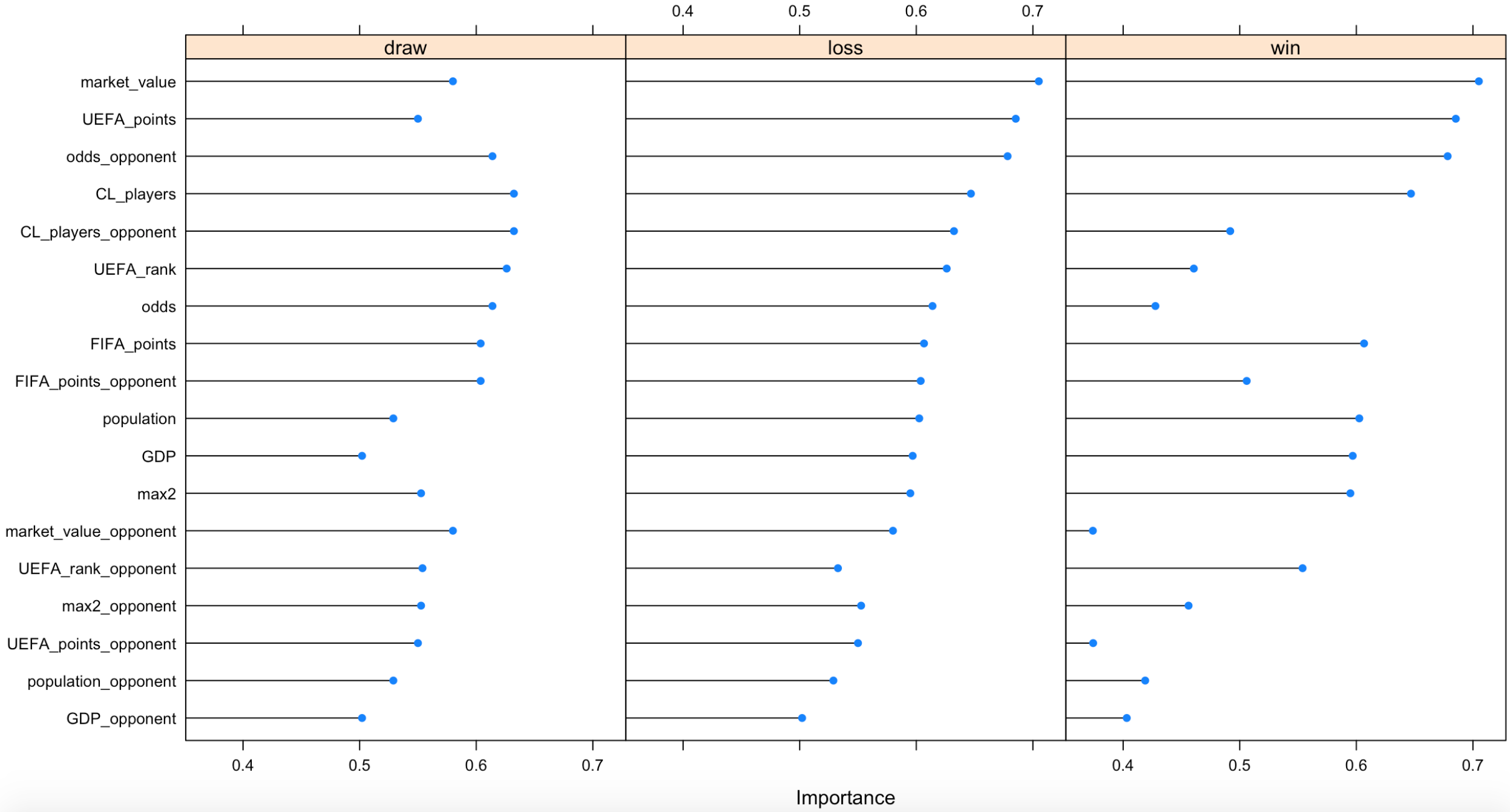
Data Preparation

Formatting, data clean-up and correction

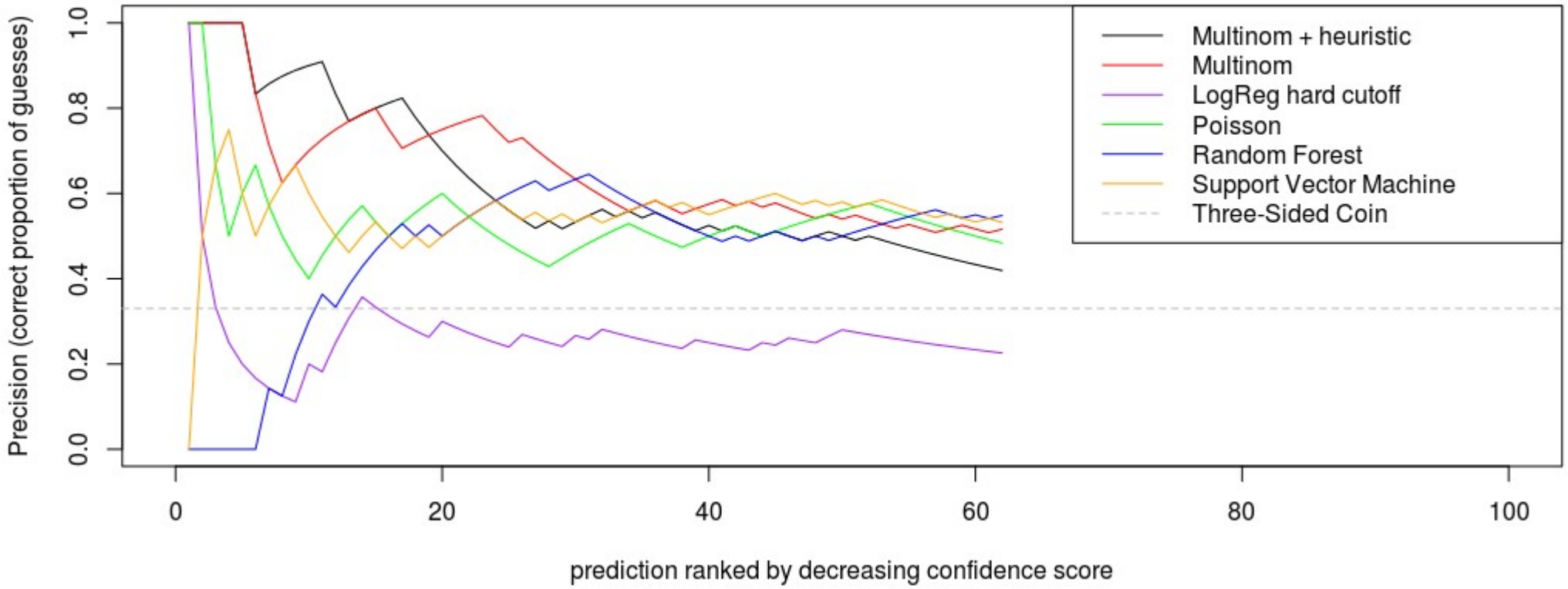
- Extending with new features
  - Nation’s GDP, population
  - Weighted historical scoring in previous editions of the tournament

  - Feature selection (LS-SVM)

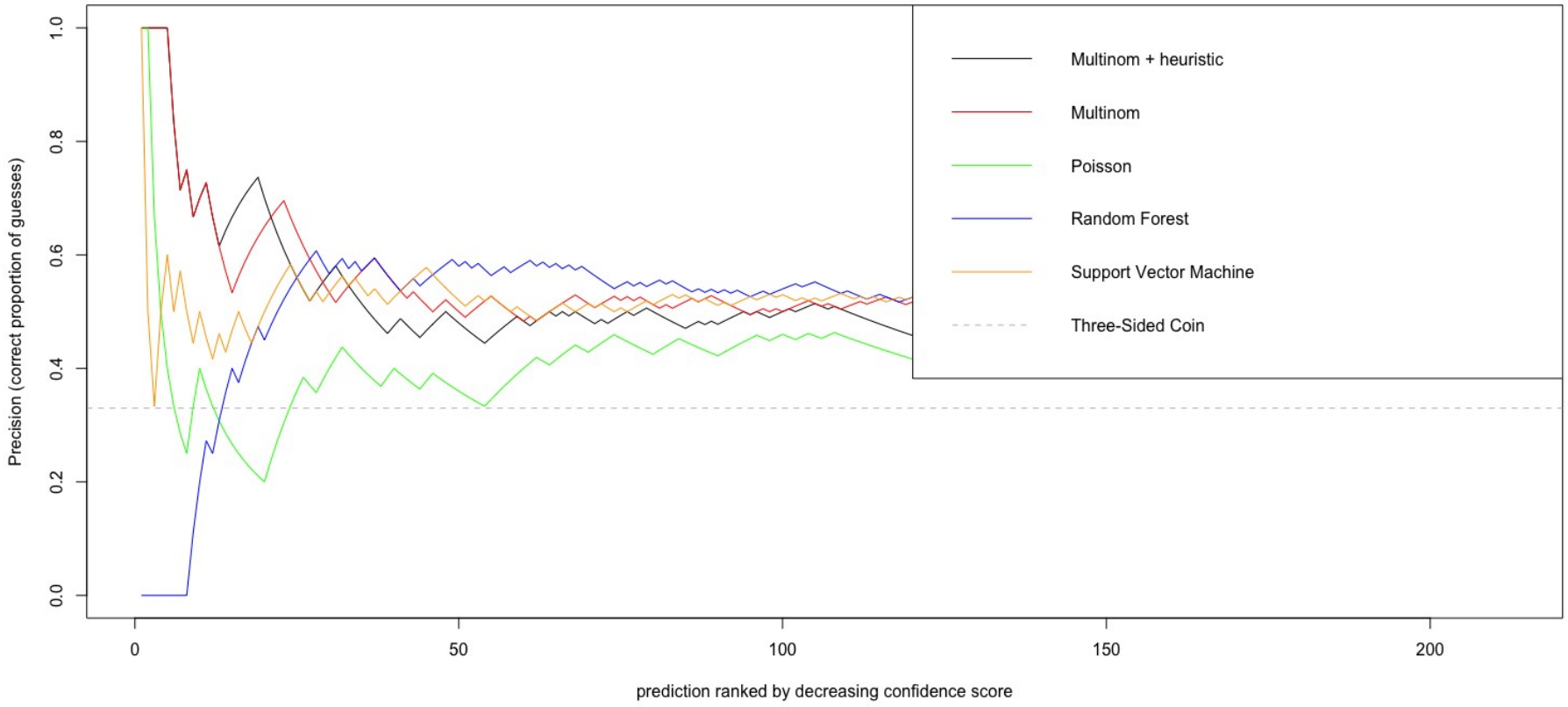
- Data visualization (t-SNE)


draw
loss
win
market_value
UEFA_points
odds_opponent
CL_players
CL_players_opponent
UEFA_rank
odds
FIFA_points
FIFA_points_opponent
population
GDP
max2
market_value_opponent
UEFA_rank_opponent
max2_opponent
UEFA_points_opponent
population_opponent
GDP_opponent
0.4
0.5
0.6
0.7
Importance

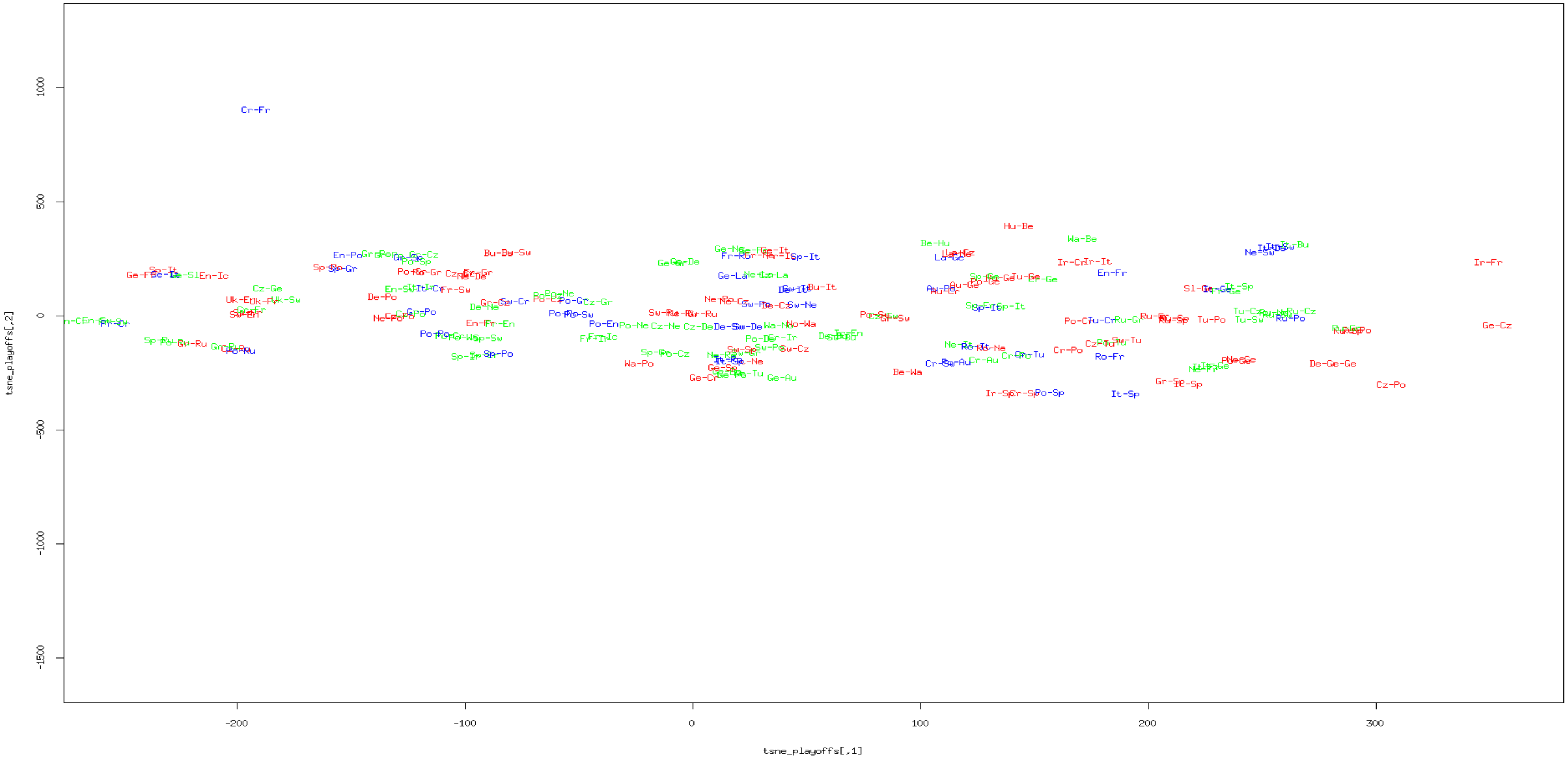
Evaluation and Model Selection



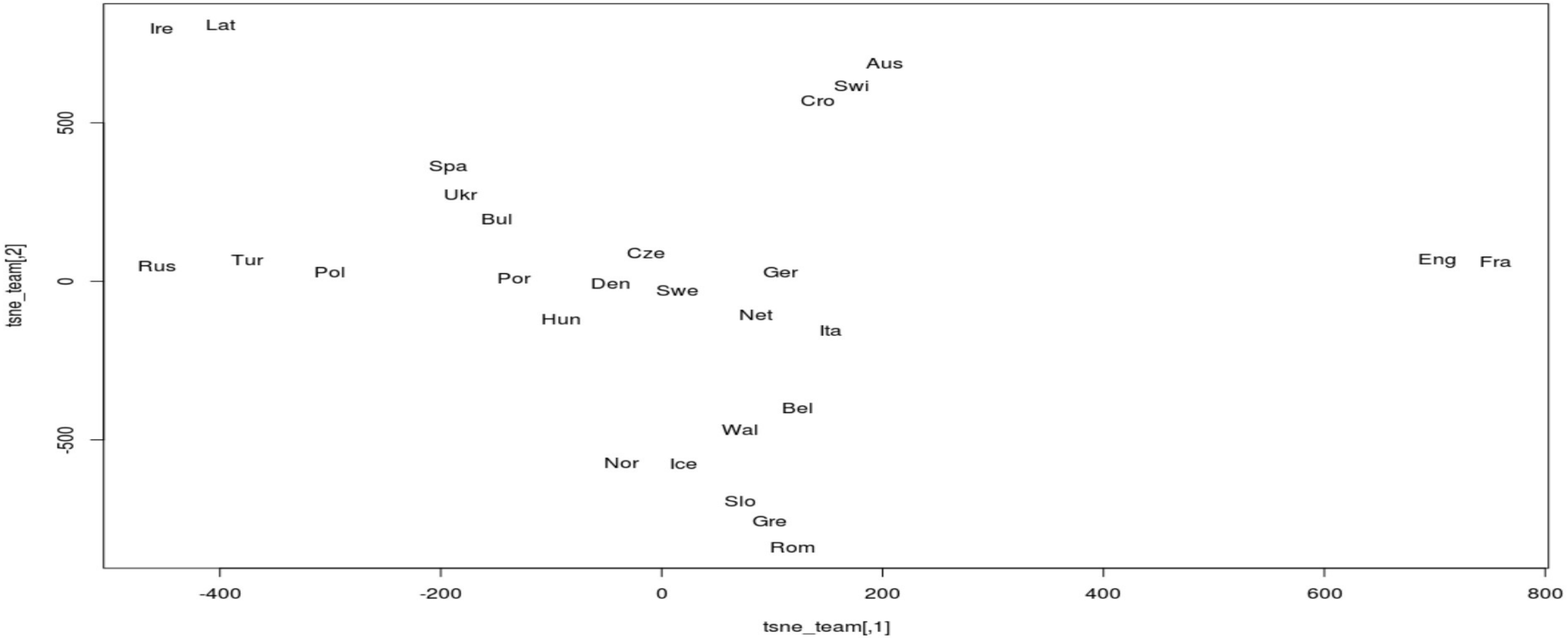
Evaluation and Model Selection



t-SNE on all games



t-SNE on team data only



Conclusions

- Probabilistic or not, models are similar in performance
- Modeling with Poisson, without additional features, throws away way to much information
- The lack of data decreases the confidence in our models' performance