Predicting the NCAA Basketball Tournament for Fun and Profit

Three Lessons for ML Projects

Jonathan Arfa, Data Scientist @ Magnetic
Brian Femiano, Data Engineer @ Magnetic
www.magnetic.com
What is March Madness?
For all 2,278 potential matchups in the tournament, submit the probability that team1 beats team2.

Teams judged on Log Loss of predicted probability (0-1) vs. actual outcome of game (0 or 1)
Log Loss for a Given Outcome & Prediction

Predicting 75% will get you 0.29 if team 1 wins, 1.39 otherwise
Lesson 1: Get the Best Data

• Team-level metrics aggregated from regular season games

• Ken Pomeroy’s team-level metrics (paid subscription data)

• Vegas betting odds for first-round games

• Distance traveled
Lesson 2: If Your Results Are Too Good To Be True, They’re Probably Wrong

- **Data leakage** - “the creation of unexpected additional information in the training data, allowing a model or machine learning algorithm to make unrealistically good predictions” [1]
  - your training data must represent only the knowledge that will exist when your model is run in the real world

- FULLY UNDERSTAND AND EXPLORE YOUR DATA BEFORE USING IT

Lesson 3: Separate Yourself From the Pack

• Gamble more - manually adjust predictions for a few games.
  • Most of the top (and the bottom) teams did this.

• Unique Data / Features
  • Use the network of regular season games better
    • If team A > team B > team C, then team A > team C.

• Take a Bayesian approach to predicting later games
  • If a low-ranked team wins the 1st two rounds, it has revealed itself to
    be a better team than previously thought. Shouldn’t we upgrade its
    chances of winning the next game?
# Performance

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>73</td>
<td>Scaling Lana</td>
<td>0.478858</td>
</tr>
<tr>
<td>74</td>
<td>Timothy Scharf</td>
<td>0.478894</td>
</tr>
<tr>
<td>75</td>
<td>Kaggler</td>
<td>0.478906</td>
</tr>
<tr>
<td>76</td>
<td>Colin Carroll</td>
<td>0.479334</td>
</tr>
<tr>
<td>77</td>
<td>oldSchool</td>
<td>0.479754</td>
</tr>
<tr>
<td>78</td>
<td>Steve Koch</td>
<td>0.479894</td>
</tr>
<tr>
<td>79</td>
<td>MachineEarning</td>
<td>0.479986</td>
</tr>
</tbody>
</table>

#79 / 341 teams

Winning Team: 0.439

**MachineEarning**: 0.480

Median of all teams: 0.489
