Machine Learning + NCAA Basketball

Cody Braun
2-22-19

Kaggle Data-Science Competition

Google provides some prize money, Kaggle runs the competition:

https://www.kaggle.com/c/mens-machine-learning-competition-2019

https://www.kaggle.com/c/womens-machine-learning-competition-2019

Last year there were about 1,000 teams

The Data

You're provided with:

- Play by play data
- Summarized box scores
- Player data
- Ranking data (AP, RPI, etc.)
- Geographical information
- Coaches
- Historical seeding
- Plus whatever else you can dig up

What You Predict

You have between Selection Sunday and the first game to produce probabilities for every possible matchups

- 64 teams, so 64*63 possible matchups (though most of them won't happen)
- Predict probability that Team A beats Team B
- Scored based on the log loss of your predictions- basically it's very bad to be wrong and confident

How Do You Represent a Team

- Data at many levels of granularity- play by plays down to the second, box scores for every game, players churn frequently
- So try and build a snapshot of team and their opponent at the time of a game
- Just use the most recent stats? Or the average for the past five years? Or a small MA window?
- How do you capture player-level stats?
- Does rebounding matter? Or rebounding relative to your opponent?
- Use old historical data?

Features - All 552 of ‘em

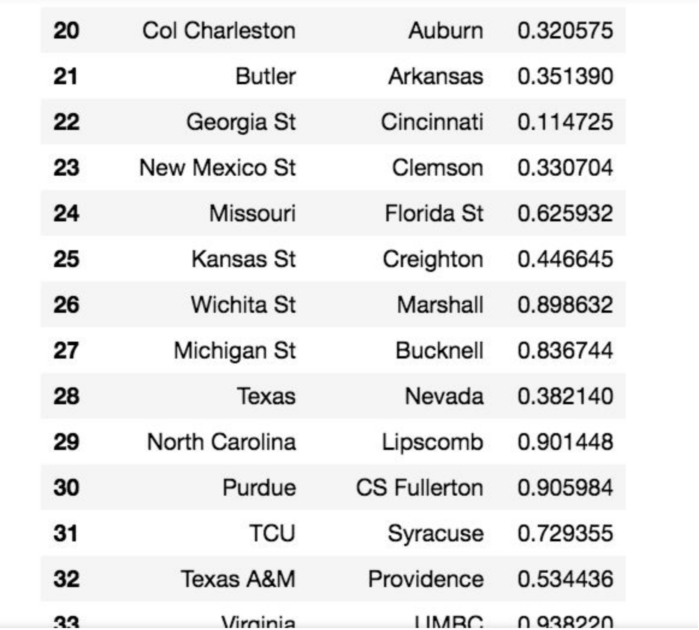
'WFGM_Perc', 'Seed', '7OT', 'ACU', 'ADE', 'AP', 'ARG', 'AUS', 'BBT', 'BCM', 'BD', 'BIH', 'BKM', 'BLS', 'BNM', 'BNT', 'BOB', 'BOW', 'BP5', 'BPI', 'BRZ', 'BUR', 'BWE', 'CJB', 'CMV', 'CNG', 'COL', 'COX', 'CPA', 'CPR', 'CRO', 'CRW', 'CTL', 'D1A', 'DAV', 'DC', 'DCI', 'DDB', 'DES', 'DII', 'DOK', 'DOL', 'DUN', 'DWH', 'EBB', 'EBP', 'ECK', 'ENT', 'ERD', 'ESR', 'FAS', 'FMG', 'FSH', 'GC', 'GRN', 'GRS', 'HAS', 'HAT', 'HER', 'HKB', 'HKS', 'HOL', 'HRN', 'IMS', 'INP', 'ISR', 'JCI', 'JEN', 'JJK', 'JNG', 'JON', 'JRT', 'KBM', 'KEL', 'KLK', 'KMV', 'KOS', 'KPI', 'KPK', 'KRA', 'LAB', 'LMC', 'LOG', 'LYD', 'LYN', 'MAS', 'MB', 'MCL', 'MGY', 'MIC', 'MKV', 'MMG', 'MOR', 'MPI', 'MSX', 'MUZ', 'MvG', 'NOL', 'NOR', 'OCT', 'OMY', 'PEQ', 'PGH', 'PH', 'PIG', 'PKL', 'PMC', 'POM', 'PPR', 'PRR', 'PTS', 'RAG', 'REI', 'REN', 'REW', 'RIS', 'RM', 'ROG', 'ROH', 'RPI', 'RSE', 'RSL', 'RT', 'RTB', 'RTH', 'RTP', 'RTR', 'SAG', 'SAP', 'SAU', 'SCR', 'SE', 'SEL', 'SFX', 'SGR', 'SIM', 'SMN', 'SMS', 'SP', 'SPR', 'SPW', 'STF', 'STH', 'STM', 'STR', 'STS', 'TBD', 'TMR', 'TOL', 'TPR', 'TRK', 'TRP', 'TRX', 'TSR', 'TW', 'UCS', 'UPS', 'USA', 'WIL', 'WLK', 'WMR', 'WMV', 'WOB', 'WOL', 'WTE', 'YAG', 'ZAM', 'Ast2', 'Blk2', 'DR2', 'FGA2', 'FGA32', 'FGM2', 'FGM32', 'FGM3_Perc2', 'FGM_Perc2', 'FTA2', 'FTM2', 'FTM_Perc2', 'OR2', 'PF2', 'Stl2', 'TO2', 'WFGM_Perc2', 'Seed2', '7OT2', 'ACU2', 'ADE2', 'AP2', 'ARG2', 'AUS2', 'BBT2', 'BCM2', 'BD2', 'BIH2', 'BKM2', 'BLS2', 'BNM2', 'BNT2', 'BOB2', 'BOW2', 'BP52', 'BPI2', 'BRZ2', 'BUR2', 'BWE2', 'CJB2', 'CMV2', 'CNG2', 'COL2', 'COX2', 'CPA2', 'CPR2', 'CRO2', 'CRW2', 'CTL2', 'D1A2', 'DAV2', 'DC2', 'DCI2', 'DDB2', 'DES2', 'DII2', 'DOK2', 'DOL2', 'DUN2', 'DWH2', 'EBB2', 'EBP2', 'ECK2', 'ENT2', 'ERD2', 'ESR2', 'FAS2', 'FMG2', 'FSH2', 'GC2', 'GRN2', 'GRS2', 'HAS2', 'HAT2', 'HER2', 'HKB2', 'HKS2', 'HOL2', 'HRN2', 'IMS2', 'INP2', 'ISR2', 'JCI2', 'JEN2', 'JJK2', 'JNG2', 'JON2', 'JRT2', 'KBM2', 'KEL2', 'KLK2', 'KMV2', 'KOS2', 'KPI2', 'KPK2', 'KRA2', 'LAB2', 'LMC2', 'LOG2', 'LYD2', 'LYN2', 'MAS2', 'MB2', 'MCL2', 'MGY2', 'MIC2', 'MKV2', 'MMG2', 'MOR2', 'MPI2', 'MSX2', 'MUZ2', 'MvG2', 'NOL2', 'NOR2', 'OCT2', 'OMY2', 'PEQ2', 'PGH2', 'PH2', 'PIG2', 'PKL2', 'PMC2', 'POM2', 'PPR2', 'PRR2', 'PTS2', 'RAG2', 'REI2', 'REN2', 'REW2', 'RIS2', 'RM2', 'ROG2', 'ROH2', 'RPI2', 'RSE2', 'RSL2', 'RT2', 'RTB2', 'RTH2', 'RTP2', 'RTR2', 'SAG2', 'SAP2', 'SAU2', 'SCR2', 'SE2', 'SEL2', 'SFX2', 'SGR2', 'SIM2', 'SMN2', 'SMS2', 'SP2', 'SPR2', 'SPW2', 'STF2', 'STH2', 'STM2', 'STR2', 'STS2', 'TBD2', 'TMR2', 'TOL2', 'TPR2', 'TRK2', 'TRP2', 'TRX2', 'TSR2', 'TW2', 'UCS2', 'UPS2', 'USA2', 'WIL2', 'WLK2', 'WMR2', 'WMV2', 'WOB2', 'WOL2', 'WTE2', 'YAG2', 'ZAM2', 'indexdiff', 'Astdiff', 'Blkdiff', 'DRdiff', 'FGAdiff', 'FGA3diff', 'FGMdiff', 'FGM3diff', 'FGM3_Percdiff', 'FGM_Percdiff', 'FTAdiff', 'FTMdiff', 'FTM_Percdiff', 'ORdiff', 'PFdiff', 'Stldiff', 'TOdiff', 'WFGM_Percdiff', 'Seeddiff', '7OTdiff', 'ACUdiff', 'ADEdiff', 'APdiff', 'ARGdiff', 'AUSdiff', 'BBTdiff', 'BCMdiff', 'BDdiff', 'BIHdiff', 'BKMdiff', 'BLSdiff', 'BNMdiff', 'BNTdiff', 'BOBdiff', 'BOWdiff', 'BP5diff', 'BPIdiff', 'BRZdiff', 'BURdiff', 'BWEdiff', 'CJBdiff', 'CMVdiff', 'CNGdiff', 'COLdiff', 'COXdiff', 'CPAdiff', 'CPRdiff', 'CROdiff', 'CRWdiff', 'CTLdiff', 'D1Adiff', 'DAVdiff', 'DCdiff', 'DCIdiff', 'DDBdiff', 'DESdiff', 'DIIdiff', 'DOKdiff', 'DOLdiff', 'DUNdiff', 'DWHdiff', 'EBBdiff', 'EBPdiff', 'ECKdiff', 'ENTdiff', 'ERDdiff', 'ESRdiff', 'FASdiff', 'FMGdiff', 'FSHdiff', 'GCdiff', 'GRNdiff', 'GRSdiff', 'HASdiff', 'HATdiff', 'HERdiff', 'HKBdiff', 'HKSdiff', 'HOLdiff', 'HRNdiff', 'IMSdiff', 'INPdiff', 'ISRdiff', 'JCIdiff', 'JENdiff', 'JJKdiff', 'JNGdiff', 'JONdiff', 'JRTdiff', 'KBMdiff', 'KELdiff', 'KLKdiff', 'KMVdiff', 'KOSdiff', 'KPIdiff', 'KPKdiff', 'KRAdiff', 'LABdiff', 'LMCdiff', 'LOGdiff', 'LYDdiff', 'LYNdiff', 'MASdiff', 'MBdiff', 'MCLdiff', 'MGYdiff', 'MICdiff', 'MKVdiff', 'MMGdiff', 'MORdiff', 'MPIdiff', 'MSXdiff', 'MUZdiff', 'MvGdiff', 'NOLdiff', 'NORdiff', 'OCTdiff', 'OMYdiff', 'PEQdiff', 'PGHdiff', 'PHdiff', 'PIGdiff', 'PKLdiff', 'PMCdiff', 'POMdiff', 'PPRdiff', 'PRRdiff', 'PTSdiff', 'RAGdiff', 'REIdiff', 'RENdiff', 'REWdiff', 'RISdiff', 'RMdiff', 'ROGdiff', 'ROHdiff', 'RPIdiff', 'RSEdiff', 'RSLdiff', 'RTdiff', 'RTBdiff', 'RTHdiff', 'RTPdiff', 'RTRdiff', 'SAGdiff', 'SAPdiff', 'SAUdiff', 'SCRdiff', 'SEdiff', 'SELdiff', 'SFXdiff', 'SGRdiff', 'SIMdiff', 'SMNdiff', 'SMSdiff', 'SPdiff', 'SPRdiff', 'SPWdiff', 'STFdiff', 'STHdiff', 'STMdiff', 'STRdiff', 'STSdiff', 'TBDdiff', 'TMRdiff', 'TOLdiff', 'TPRdiff', 'TRKdiff', 'TRPdiff', 'TRXdiff', 'TSRdiff', 'TWdiff', 'UCSdiff', 'UPSdiff', 'USAdiff', 'WILdiff', 'WLKdiff', 'WMRdiff', 'WMVdiff', 'WOBdiff', 'WOLdiff', 'WTEdiff', 'YAGdiff', 'ZAMdiff'

The Model

- There aren’t that many games so probably want something that generalizes well
- Could treat it as a classification problem and just try and predict winner
- Or try and predict point differential and map that to win probabilities
- Have a ton of features, so need a model that handles this well
- Or do some feature selection
- Ended up doing some dimensionality reduction and a gradient boosting classifier

Validation

- Just look at how it performs against past tournaments (regular games may have a different pattern)
- It gets pretty close to 90%
- The most useful features are intuitive:
    - Playing at home
    - FGM diff
    - Seed diff
- Many of the ratings agencies are terrible



20	Col Charleston	Auburn	0.320575
21	Butler	Arkansas	0.351390
22	Georgia St	Cincinnati	0.114725
23	New Mexico St	Clemson	0.330704
24	Missouri	Florida St	0.625932
25	Kansas St	Creighton	0.446645
26	Wichita St	Marshall	0.898632
27	Michigan St	Bucknell	0.836744
28	Texas	Nevada	0.382140
29	North Carolina	Lipscomb	0.901448
30	Purdue	CS Fullerton	0.905984
31	TCU	Syracuse	0.729355
32	Texas A&M	Providence	0.534436
33	Virginia	UMBC	0.938220