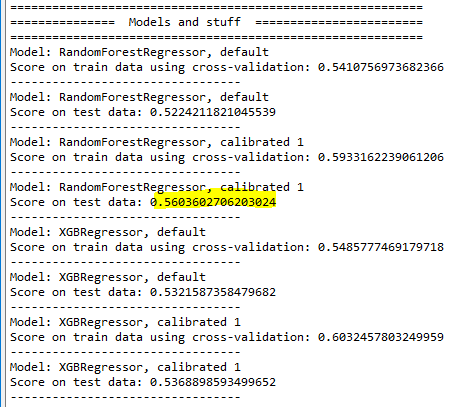
Individually optimal parameters for RandomForestRegressor:

* RandomForestRegressor calibrated 1 – so far the best (0.557 on test). XGBoost calibrated 2 – 0.54 on test data)
* Number of trees – 70 => 0.55 R^2. Latest => 700
* Criterion – mse
* Max\_features – auto
* Tree\_depth – no limit. Latest => 30
* Min\_samples\_leaf = 2. Latest => 3
* Max\_leaves\_nodes -> max
* bootstrap => definately True (default)
* verbose => 0

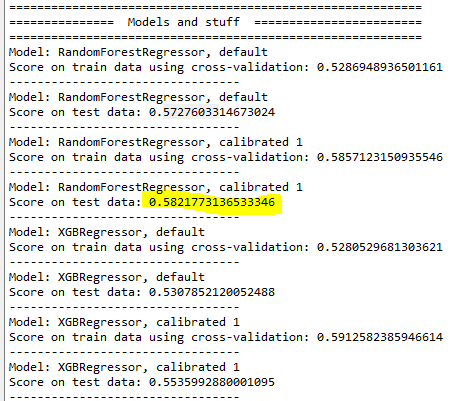
overall, default model (with no parameters calibrated) gives 0.52 on test data. The calibrated model (where number of trees = 70 and min\_samples\_leaf = 2 gives 0.63. On another try the improvement was much more modest – from 0.555 to 0.568

Individually optimal parameters for XGBoost:

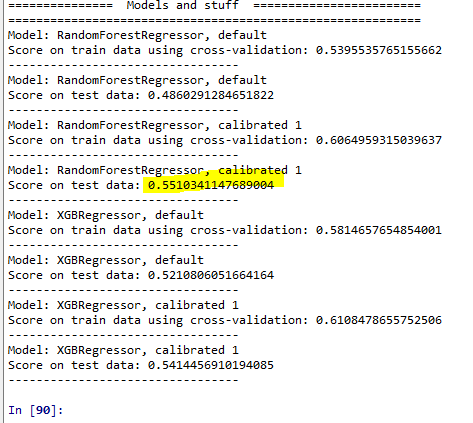
* max\_depth=5 from grid search. 7 from manual (or 4 if more conservative) – check with test data! Range =
* Min\_child\_weight = 9 from grid search. 7 from simple loop. Range – [1-11].
* Gamma = 1. Range = [0 – 20]. n/s. Can try 0
* subsample = 0.8 by grid search. 0.7 by loop. Range – [0.5-1]
* colsample\_bytree = 0.7 by grid search. 0.8 by loop. range = [0.6-1]
* colsample\_bylevel = 1 for grid search. 0.65 by loop. range = [0.3 – 1]
* 'reg\_alpha': 0.01. range=[0 – 0.015]
* Reg\_lambda = 1
* Scale\_pos\_weight = 1 range=[0-1000]
* Learning\_rate = 0.14 by grid search; 0.18 by loop. Range = [0.10 – 0.30]. 0.1 for final
* n\_estimators = 175 with grid search; 199 by loop. Range = 150-220
* {'learning\_rate': 0.075, 'n\_estimators': 275},
* n\_trhead = 4. Affects only speed. On my laptop 8 would be better, but for should be fine.
* Max\_delta\_step=0. Range – [0 – 100]
* model = XGBRegressor(max\_depth=5, min\_child\_weight=9, n\_estimators=275,
* subsample=0.8, colsample\_bytree=0.7, colsample\_bylevel=1,
* gamma=1, reg\_alpha=0.01, reg\_lambda=1, learning\_rate=0.075, nthread=8)
* With court\_dummies and randoms seeds = 1



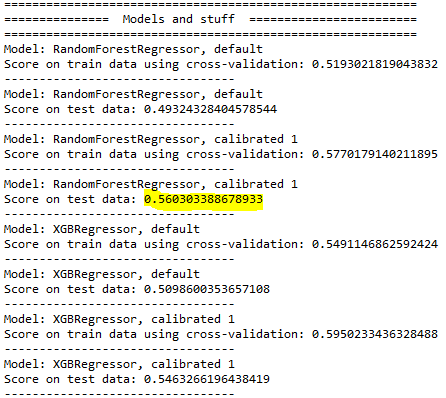
* Without court\_dummies and randoms seeds = 1



* With court\_dummies and randoms seeds = 2

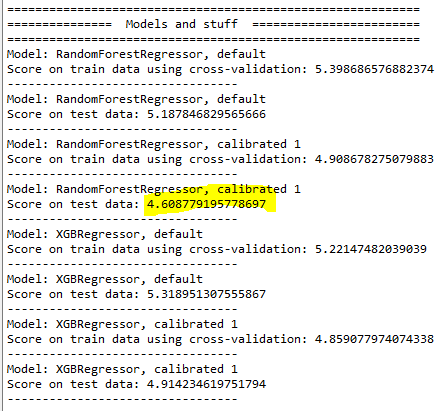


* Without court\_dummies and randoms seeds = 2



Best model with MSE as a scoring method

(with start\_date\_float, but without month\_float\_start\_date and month\_int\_start\_date)



Exclude

keep

'start\_date\_day'

'start\_date\_month'

('start\_date\_month\_float'

'start\_date\_float'

'start\_date\_quarter'

'start\_date\_year'

'court\_name'

'court\_id'

'duration\_m',

'region',

# 'reciept\_procedure\_case\_join',

# 'reciept\_procedure\_case\_split',

# 'reciept\_procedure\_court\_belonging',

# 'reciept\_procedure\_first',

# 'reciept\_procedure\_quick',

# 'reciept\_procedure\_ruling\_cancel',

# 'money\_amount\_req\_less\_2',

# 'money\_amount\_req\_zero',

# 'p\_to\_p\_dummy',

# 'p\_to\_b\_dummy',

# 'b\_to\_b\_dummy',

# 'b\_to\_p\_dummy',

# 'b\_and\_p\_other\_dummy',

# 'case\_matter\_contract\_breach',

# 'case\_matter\_insurance',

# 'case\_matter\_loan\_recovery',

# 'case\_matter\_loss\_recovery',

# 'case\_matter\_money\_recovery',

# 'case\_matter\_other',

# 'case\_matter\_pay\_recovery',

# 'case\_matter\_property',

# 'case\_matter\_rent',

# 'case\_matter\_support',

# 'court\_aizkraukles\_rajona\_tiesa',

# 'court\_alūksnes\_rajona\_tiesa',

# 'court\_bauskas\_rajona\_tiesa',

# 'court\_cēsu\_rajona\_tiesa',

# 'court\_daugavpils\_tiesa',

# 'court\_dobeles\_rajona\_tiesa',

# 'court\_gulbenes\_rajona\_tiesa',

# 'court\_jelgavas\_tiesa',

# 'court\_jēkabpils\_rajona\_tiesa',

# 'court\_kuldīgas\_rajona\_tiesa',

# 'court\_kurzemes\_rajona\_tiesa',

# 'court\_liepājas\_tiesa',

# 'court\_limbažu\_rajona\_tiesa',

# 'court\_madonas\_rajona\_tiesa',

# 'court\_ogres\_rajona\_tiesa',

# 'court\_rēzeknes\_tiesa',

# 'court\_rīgas\_pilsētas\_kurzemes\_rajona\_tiesa',

# 'court\_rīgas\_pilsētas\_latgales\_priekšpilsētas\_tiesa',

# 'court\_rīgas\_pilsētas\_pārdaugavas\_tiesa',

# 'court\_rīgas\_pilsētas\_vidzemes\_priekšpilsētas\_tiesa',

# 'court\_rīgas\_pilsētas\_zemgales\_priekšpilsētas\_tiesa',

# 'court\_rīgas\_pilsētas\_ziemeļu\_rajona\_tiesa',

# 'court\_rīgas\_rajona\_tiesa',

# 'court\_rīgas\_rajona\_tiesa\_jūrmalas\_tiesu\_nams',

# 'court\_saldus\_rajona\_tiesa',

# 'court\_talsu\_rajona\_tiesa',

# 'court\_tukuma\_rajona\_tiesa',

# 'court\_valkas\_rajona\_tiesa',

# 'court\_valmieras\_rajona\_tiesa',

# 'court\_ventspils\_tiesa',

'loadiness\_of\_court',

# 'not\_subject\_to\_duty\_not\_zero',

# 'persons\_and\_compnies\_answering',

# 'lives\_abroad\_over\_persons\_and\_companies\_answering',

# 'start\_date\_month',

# 'start\_date\_quarter',

# 'persons\_and\_compnies\_started',

# 'single\_person\_or\_company\_started',

# 'single\_person\_or\_company\_answered',

# 'court\_judge\_count',

# 'court\_productivity',

# 'money\_amount\_requested',

# 'person\_answering\_to\_case',

# 'legal\_entity\_answering\_to\_case',

# 'person\_started\_the\_case',

# 'legal\_entity\_started\_the\_case',

# 'court\_meetings\_set', --------------- affects some 20%

# 'court\_meetings\_happened',

# 'admin\_penalty', < - fix or kick out

‘admin\_penalty\_not\_zero’

# 'lives\_abroad',

# 'not\_subject\_to\_duty',

'money\_amount\_from\_2\_to\_15',

'money\_amount\_from\_15\_to\_50',

'money\_amount\_from\_50\_to\_500',

'money\_amount\_from\_500\_to\_1000',

'money\_amount\_from\_1000\_to\_2000',

'money\_amount\_from\_2000\_to\_5000',

'money\_amount\_from\_5000\_to\_50\_000',

'money\_amount\_from\_50\_000\_to\_1000\_000',

'money\_amount\_over\_1000\_000',

# 'court\_meetings\_set\_is\_zero',

# 'court\_meetings\_happened\_is\_zero',

Best without time variables:

