GROUP 3's TOP 10

FEATURES FOR PREDICTING BASEBALL SALARIES

10. Cumulative Isolated power `['ISO_cum']`
(Total bases - Hits / At Bats)

FEATURES FOR PREDICTING BASEBALL SALARIES

9. Pitching Saves `['pSV']`

10. Cumulative Isolated power `['ISO_cum']`
(Total bases - Hits / At Bats)

FEATURES FOR PREDICTING BASEBALL SALARIES

8. Cumulative Plate Appearances `['PA_cum']`
At bats + Bases on balls + Hit by pitcher + Sacrifice Hits+ Sacrifice Fliers + Times on Defensive Interference

9. Pitching Saves `['pSV']`

10. Cumulative Isolated power `['ISO_cum']`
(Total bases - Hits / At Bats)

FEATURES FOR PREDICTING BASEBALL SALARIES

7. Cumulative Home runs `['HR_cum']`

8. Cumulative Plate Appearances `['PA_cum']`
At bats + Bases on balls + Hit by pitcher + Sacrifice Hits+ Sacrifice Fliers + Times on Defensive Interference

9. Pitching Saves `['pSV']`

10. Cumulative Isolated power `['ISO_cum']`
(Total bases - Hits / At Bats)

FEATURES FOR PREDICTING BASEBALL SALARIES

6. Cumulative Pitching Wins `['pW_cum']`

7. Cumulative Home runs `['HR_cum']`

8. Cumulative Plate Appearances `['PA_cum']`
At bats + Bases on balls + Hit by pitcher + Sacrifice Hits+ Sacrifice Fliers + Times on Defensive Interference

9. Pitching Saves `['pSV']`

10. Cumulative Isolated power `['ISO_cum']`
(Total bases - Hits / At Bats)

FEATURES FOR PREDICTING BASEBALL SALARIES

5. Age ['age']

FEATURES FOR PREDICTING BASEBALL SALARIES

4. **Cumulative Games** `['G_cum']`
(Sum of career games)

5. **Age** `['age']`

FEATURES FOR PREDICTING BASEBALL SALARIES

3. Log of years with club `['log_ywc']`
(Consecutive years with same club)

4. Cumulative Games `['G_cum']`
(Sum of career games)

5. Age `['age']`

FEATURES FOR PREDICTING BASEBALL SALARIES

2. Year `['yearID']`

3. Log of years with club `['log_ywc']`
(Consecutive years with same club)

4. Cumulative Games `['G_cum']`
(Sum of career games)

5. Age `['age']`

FEATURES FOR PREDICTING BASEBALL SALARIES

1. Cumulative Fielding Average `['fFA_cum']`
(Putouts + Assists)/(Putouts + Assists + Errors)

General comments

- (log of) Cumulative stats were much better features
- Common defensive and offensive meta statistics useful http://www.baseball-almanac.com/
- Importing data from pitching and fielding useful
- ‘Years with club’ captures some of the switching / resigning salary variance
- Predicting on log of salary with degree 3 polynomial

Batters vs. Pitchers

- Where batting data didn't exist, often pitching data did and vice-versa
- Set missing data to zero IF the other type exists
- Remove remaining NaNs
- Results in only a few hundred dropped records

Selecting best features

```
['birthMonth', 'nameNick', 'weight', 'height', 'bats', 'throws', 'yearID',
'teamID', 'G', 'G_batting', 'AB', 'R', 'H', 'X2B', 'X3B', 'HR', 'RBI', 'SB','CS',
'BB', 'SO', 'IBB', 'HBP', 'SH', 'SF', 'GIDP', 'pW', 'pL', 'pG', 'pGS', 'pCG',
'pSHO', 'pSV', 'pIPouts', 'pH', 'pER', 'pHR', 'pBB', 'pSO', 'pERA', 'pWP',
'pHBP', 'pBK', 'pBFP', 'pGF', 'pR', 'fPO', 'fA', 'fE', 'fInnOuts', 'age',
'Batting_ratio', 'pWHIP', 'TBP', 'PA', 'TB', 'AVG', 'BOB', 'HRR', 'OBP', 'SA',
'OBPS', 'SOR', 'ISO', 'fFA', 'fRF', 'ywc', 'G_cum', 'G_batting_cum',
'AB_cum','R_cum', 'H_cum', 'X2B_cum', 'X3B_cum', 'HR_cum', 'RBI_cum', 'SB_cum',
'CS_cum', 'BB_cum', 'SO_cum', 'IBB_cum', 'HBP_cum', 'SH_cum', 'SF_cum',
'GIDP_cum', 'pW_cum', 'pL_cum', 'pG_cum', 'pGS_cum', 'pCG_cum', 'pSHO_cum',
'pSV_cum', 'pIPouts_cum', 'pH_cum', 'pER_cum', 'pHR_cum', 'pBB_cum',
'pSO_cum','pERA_cum', 'pWP_cum', 'pHBP_cum', 'pBK_cum', 'pBFP_cum', 'pGF_cum',
'pR_cum', 'fPO_cum', 'fA_cum', 'fE_cum', 'fInnOuts_cum', 'pWHIP_cum', 'fFA_cum',
'fRF_cum', 'TBP_cum', 'PA_cum', 'TB_cum', 'AVG_cum', 'BOB_cum', 'HRR_cum',
'OBP_cum', 'SA_cum', 'OBPS_cum', 'SOR_cum', 'ISO_cum', 'teamID_ranking',
'is_NYA', 'log_ywc', 'log_G', 'log_G_batting', 'log_AB', 'log_R', 'log_H',
'log_X2B', 'log_X3B', 'log_HR', 'log_RBI', 'log_SB', 'log_CS', 'log_BB',
'log_SO', 'log_IBB', 'log_HBP', 'log_SH', 'log_SF', 'log_GIDP', 'log_pW',
'log_pL', 'log_pG', 'log_pGS', 'log_pCG', 'log_pSHO','log_pSV', 'log_pIPouts',
'log_pH', 'log_pER', 'log_pHR', 'log_pBB', 'log_pSO', 'log_pERA', 'log_pWP',
'log_pHBP', 'log_pBK', 'log_pBFP', 'log_pGF', 'log_pR','log_fPO', 'log_fA',
'log_fE', 'log_fInnOuts', 'log_pWHIP', 'log_fFA', 'log_fRF','log_TBP', 'log_PA',
'log_TB', 'log_AVG', 'log_BOB', 'log_HRR', 'log_OBP', 'log_SA', 'log_OBPS',
'log_SOR', 'log_ISO']
```

Selecting best features

```
salary_cum    91%
salary    80%
fFA_cum   68%
G_cum     65%  65%
fInnOuts_cum 56%
SOR_cum   56%
fE_cum    55%
OBP_cum   54%
TBP_cum   54%
AVG_cum   54%
SA_cum    53%
OBPS_cum  52%
ISO_cum   50%
fRF_cum   50%
fA_cum    50%
age       50%
BOB_cum   48%
ywc       46%
HRR_cum   46%
fPO_cum   46%
G_batting_cum 44%
log_ywc   44%
yearID    43%
SO_cum    41%
```

Abs(Correlation) with log_salary

Selecting best features

```
['fFA_cum', 'G_cum','G_batting_cum','fInnOuts_cum',
'OBPS_cum','OBP_cum','TBP_cum',‘AVG_cum','SA_cum','
PA_cum','AB_cum','SO_cum','TB_cum','X2B_cum','H_cum
','R_cum','SF_cum','SOR_cum','BB_cum','RBI_cum','lo
g_ywc','ISO_cum','HR_cum','GIDP_cum','pSO_cum','fE_
cum','HBP_cum','BOB_cum','age','IBB_cum','X3B_cum',
'pIPouts_cum','pBFP_cum','pW_cum','yearID','pH_cum'
,'pHR_cum','pL_cum','pBB_cum','pER_cum','pR_cum','p
WP_cum','fA_cum','pG_cum','fRF_cum','pHBP_cum','CS_
cum','fPO_cum','SH_cum','pWHIP_cum','pGS_cum','SB_c
um','pERA_cum','pSHO_cum','pCG_cum','pGF_cum','pSV_
cum','pBK_cum','Batting_ratio','nameNick','weight',
'bats','height','birthMonth','throws','is_NYA','tea
mID']
```

Selecting best features

```
Features:  ('fFA_cum',)

MAPE:  104.379537106

R^2:  0.503669380356

MSE:  0.932072473917

MAE:  0.773107186223
```

Selecting best features

```
Features:  ('fFA_cum', 'yearID')

MAPE:  83.83813359

R^2:  0.618592637311

MSE:  0.716255032517

MAE:  0.645030874818
```

Selecting best features

```
Features:  ('fFA_cum', 'yearID',
'log_ywc')

MAPE:  71.1490083069

R^2:  0.682667508573

MSE:  0.595927127267

MAE:  0.583478420927
```

Selecting best features

```
Features:  ('yearID', 'G_cum', 'age',
'pW_cum')

MAPE:  57.0897741579

R^2:  0.795143211489

MSE:  0.384706012704

MAE:  0.451887290472
```

Selecting best features

```
Features:  ('yearID', 'G_cum',
'log_ywc', 'age', 'pW_cum')

MAPE:  50.9621949766

R^2:  0.815535443476

MSE:  0.346410897786

MAE:  0.431782713551
```

Selecting best features

```
Features:  ('fFA_cum', 'yearID',
'G_cum', 'log_ywc', 'age', 'pW_cum')

MAPE:  49.400104754

R^2:  0.823469850274

MSE:  0.33151066419

MAE:  0.420049270536
```

Selecting best features

```
Features:  ('fFA_cum', 'yearID',
'G_cum', 'log_ywc', 'age', 'HR_cum',
'pW_cum')

MAPE:  48.3102080717

R^2:  0.83219409585

MSE:  0.315127171342

MAE:  0.408123503634
```

Selecting best features

```
Features:  ('fFA_cum', 'yearID',
'PA_cum', 'log_ywc', 'age', 'HR_cum',
'pSV_cum', 'pW_cum')

MAPE:  47.028007769

R^2:  0.841751365826

MSE:  0.297179319814

MAE:  0.395461909081
```

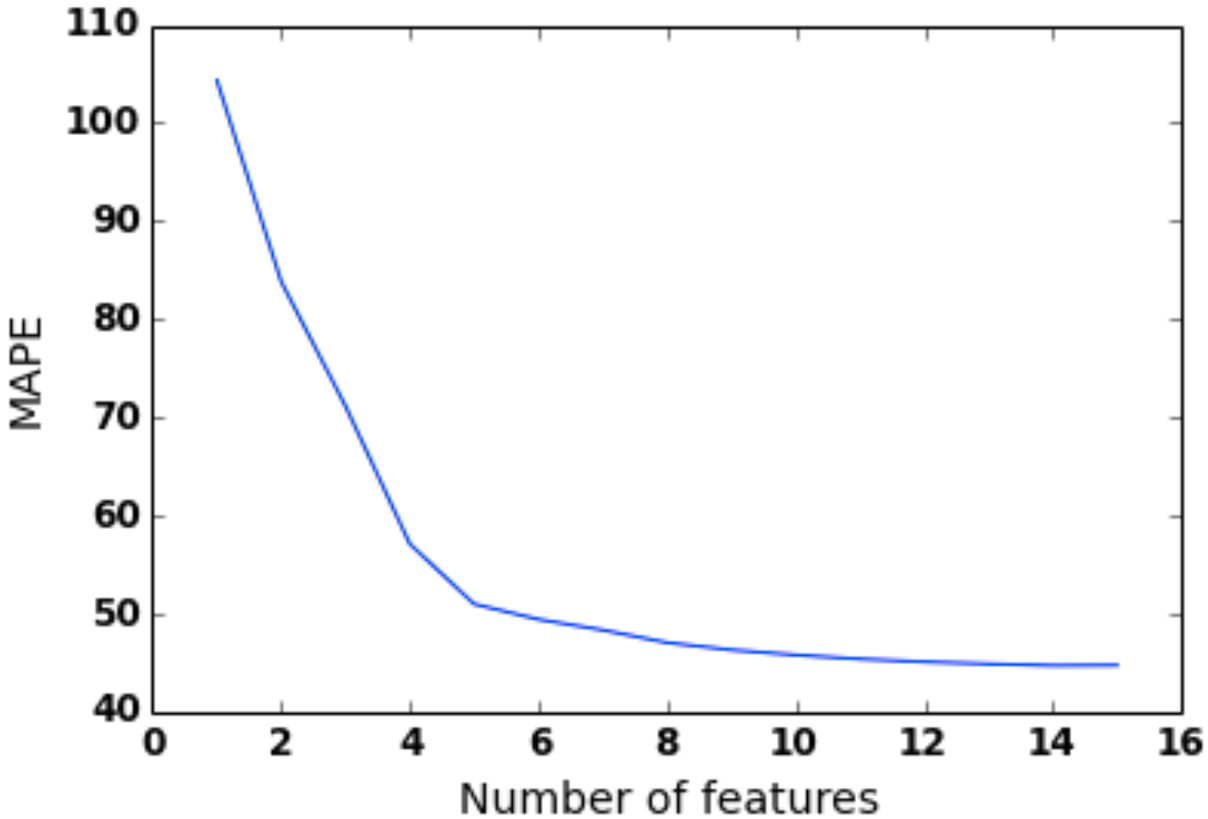
Selecting best features

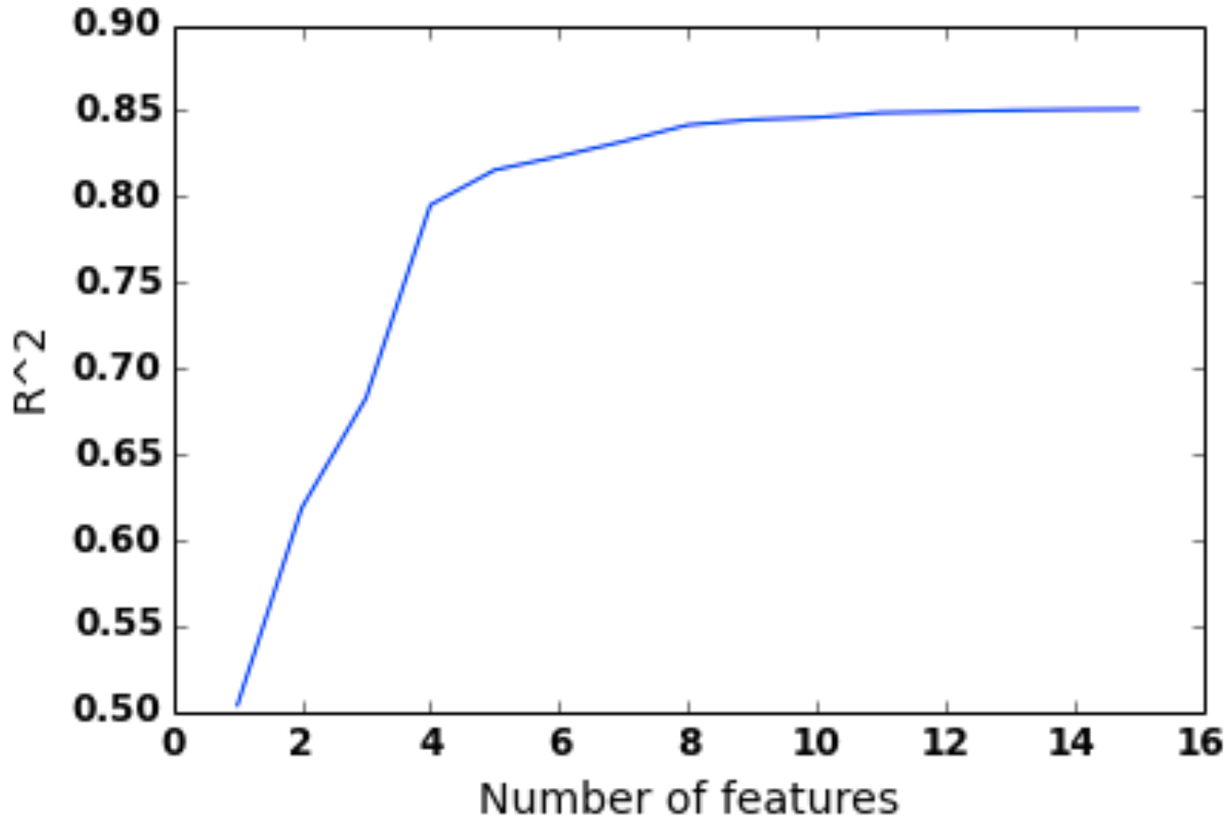
```
Features:  ('fFA_cum', 'yearID',
'G_cum', 'PA_cum', 'log_ywc', 'age',
'HR_cum', 'pSV_cum', 'pW_cum')

MAPE:  46.284537001

R^2:  0.844870530072

MSE:  0.291321758302

MAE:  0.391105291524
```

Selecting best features

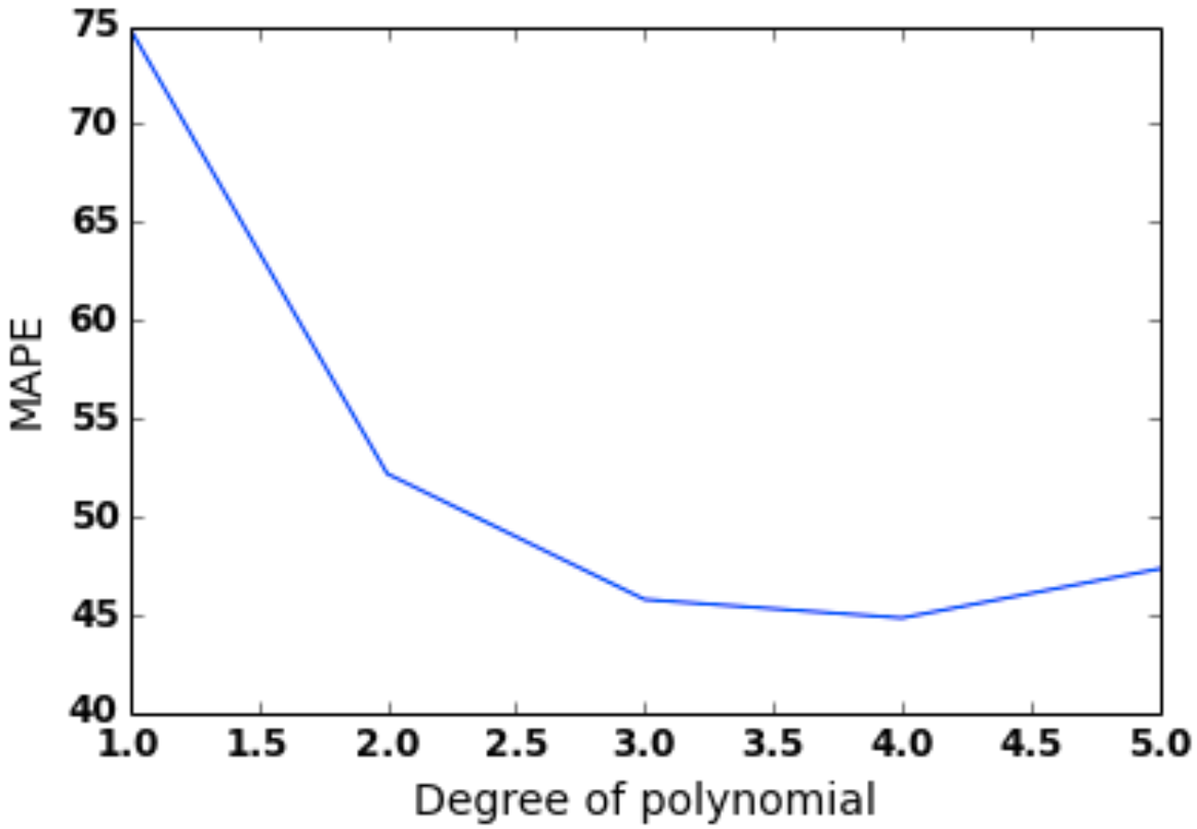
```
Features:  ('fFA_cum', 'yearID', 'G_cum',
'PA_cum', 'log_ywc', 'age', 'HR_cum',
'pSV_cum', 'ISO_cum', 'pW_cum')

MAPE:  45.7685332902

R^2:  0.846141145698

MSE:  0.288935635416

MAE:  0.387752166797
```

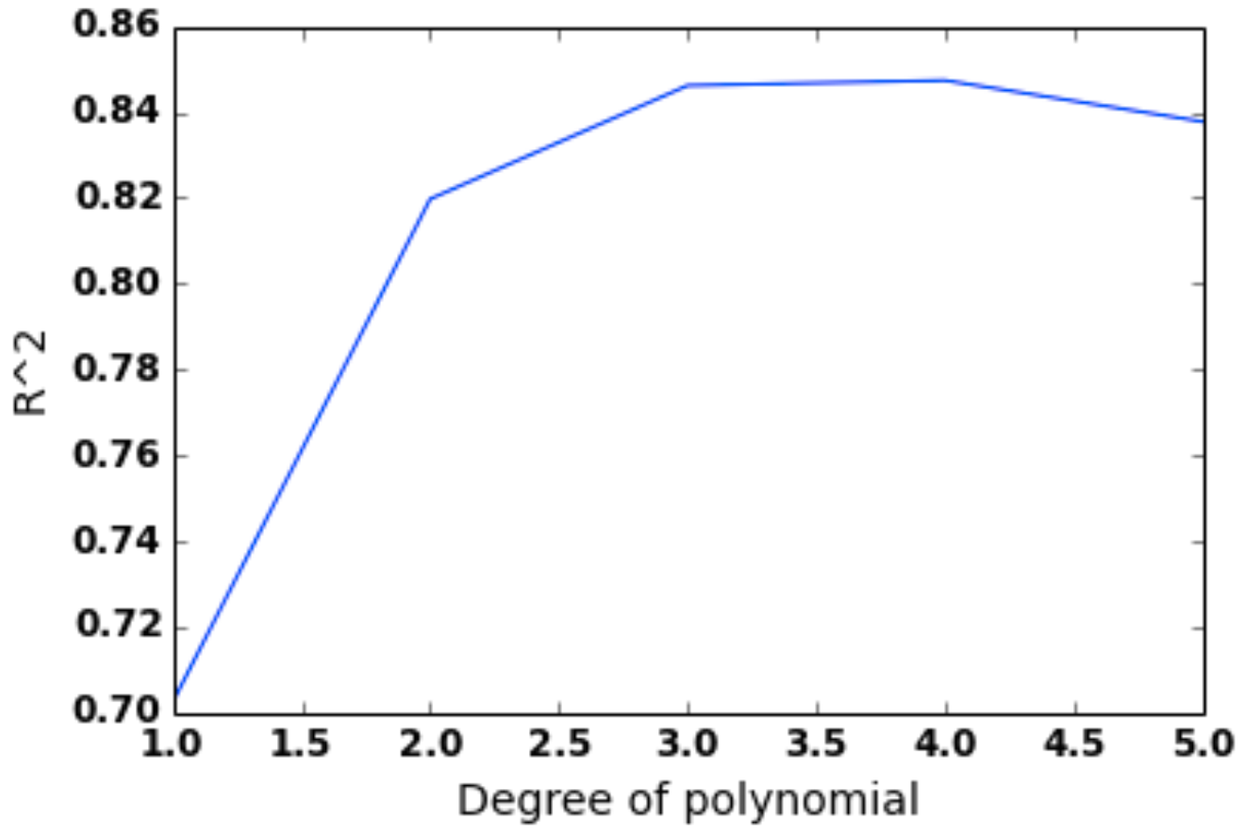
Over-fitting

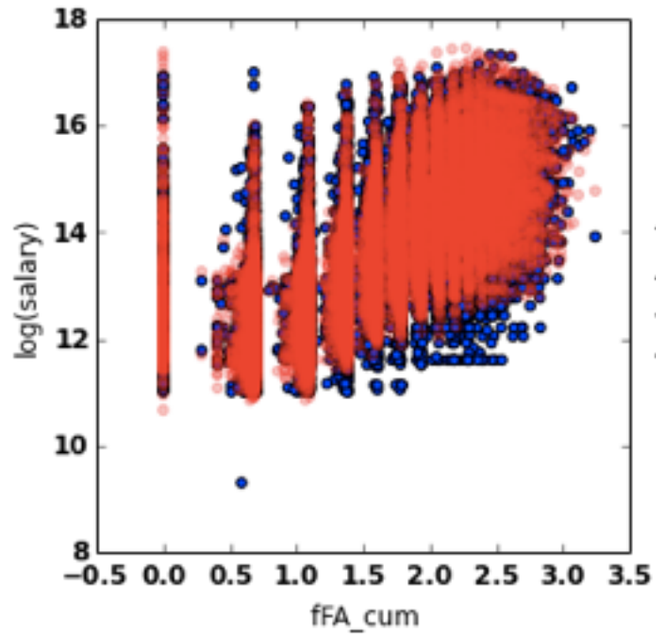


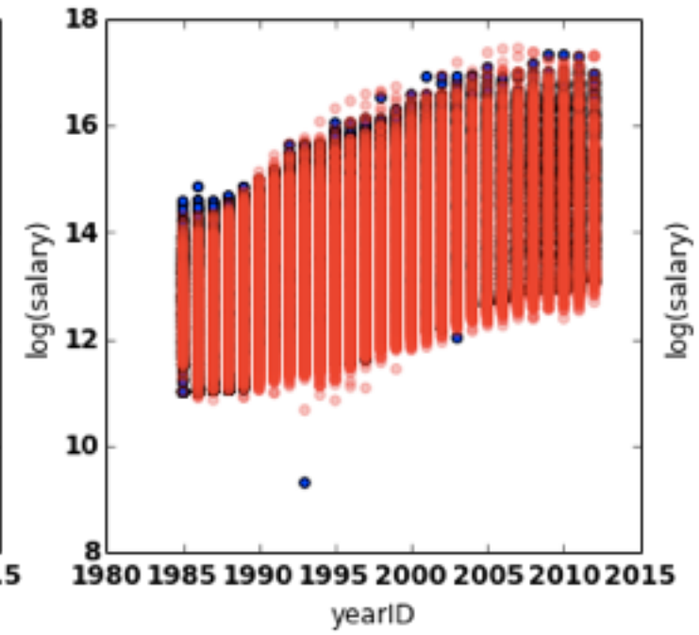


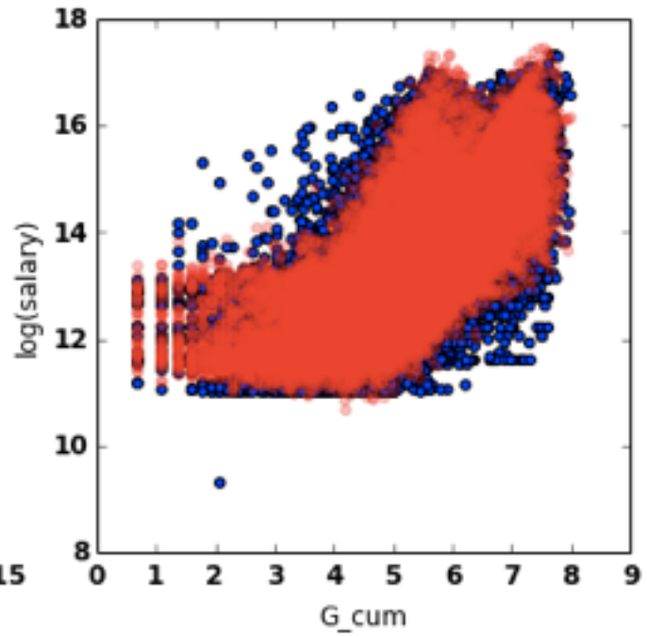
Over-fitting

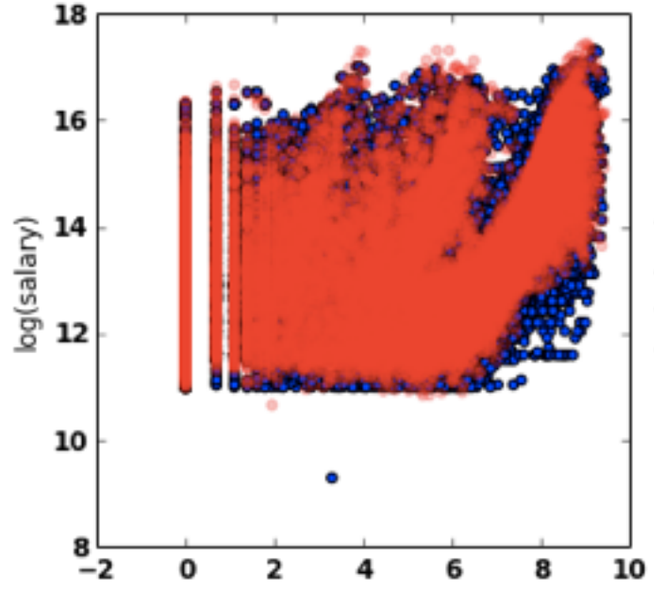


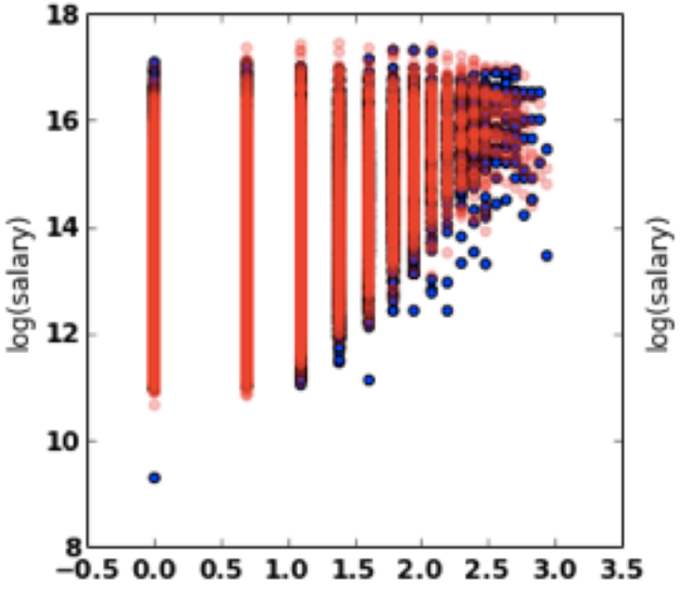


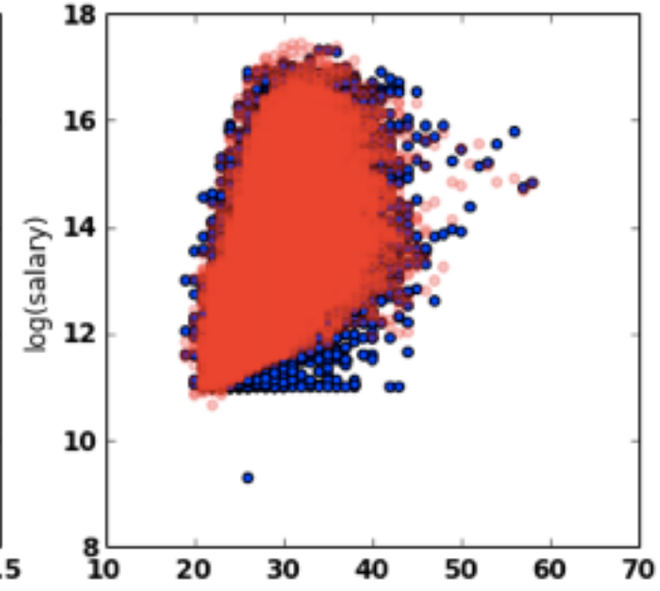


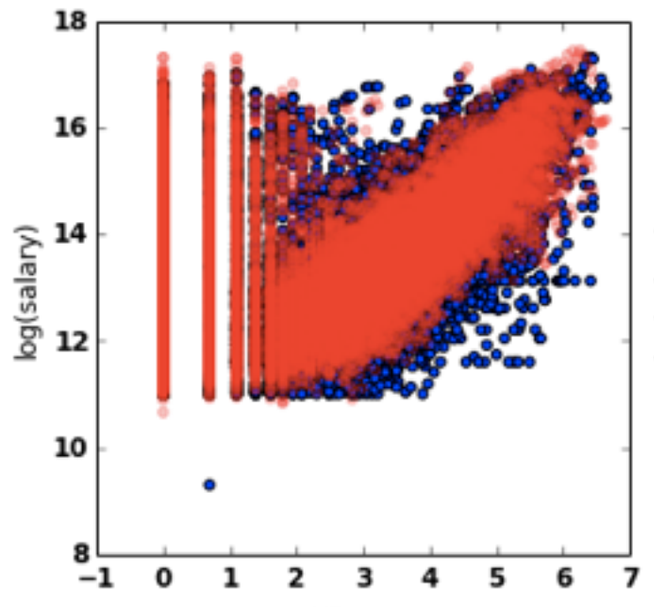


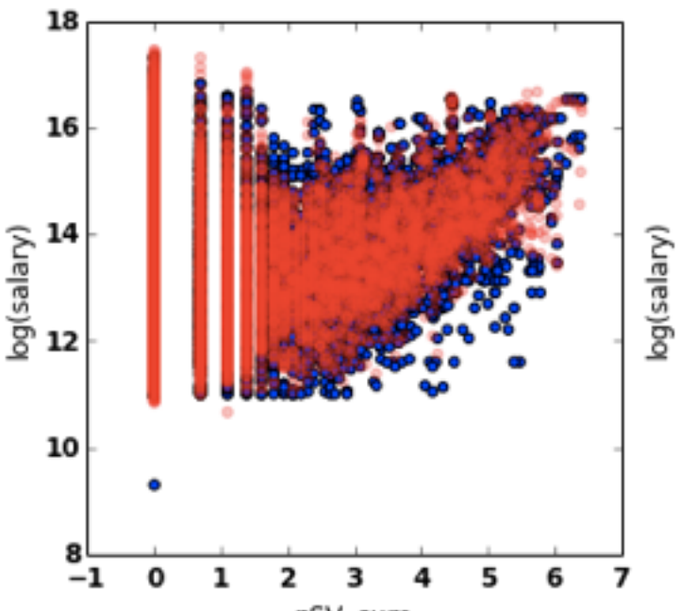


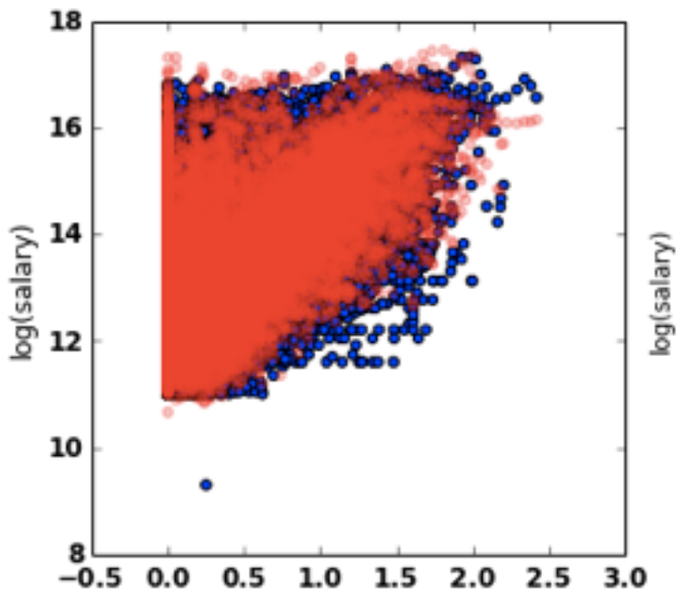


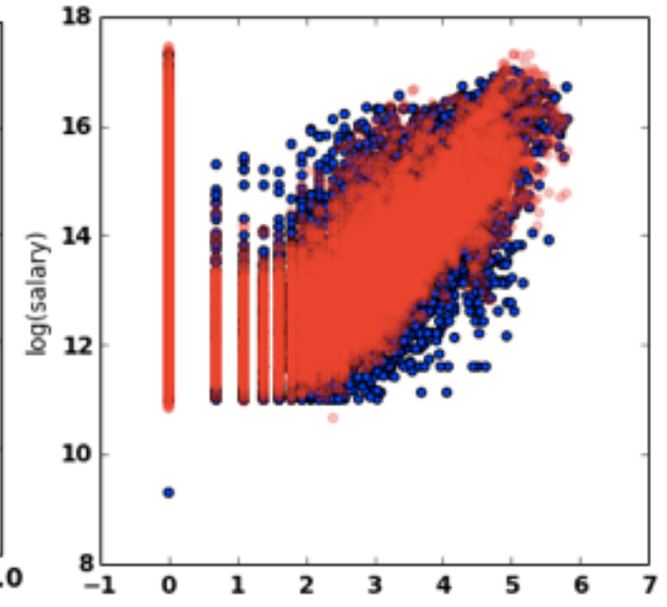












Features:
'fFA_cum'
'yearID'
'G_cum'
'PA_cum'
'log_ywc'
'age'
'HR_cum'
'pSV_cum'
'ISO_cum'
'pW_cum'

MAPE: 45.77
R^2: 0.8461
MSE: 0.2889
MAE: 0.3878