NCAA Data in Spark – Henry Tessier

Motivation

- The dataset I used contained team-level data for the 2018 NCAA basketball season.
- I was interested in how a team's Win/Loss ratio was impacted by their strength of schedule, efficiency margin, and poll ranking.
- Also interested in differences in average team strength between conferences

Team	Rank	Actual_Seed	Wins	Losses	AdjEM	SOS_Pyth	Conference
Virginia	1	1	31	2	32.15	9.99	ACC
Villanova	2	1	30	4	31.41	10.23	BE
Duke	3	2	26	7	29.13	10.9	ACC
Purdue	4	2	28	6	26.67	8.55	B10
North_Carolina	5	2	25	10	25.03	14.05	ACC
Michigan_St	6	3	29	4	26.35	6.5	B10

Code Snippets

- Created a new column of win percentage to use in regression

```
data = data.withColumn('WinPct', data.Wins / (data.Wins + data.Losses))
```

- Queried to get grouped conference average efficiencies

```
Conf = sqlc.sql(""" SELECT Conference, Round(Mean(AdjEM),2) as Rating FROM ncaa Group By Conference Order By Rating DESC """)
```

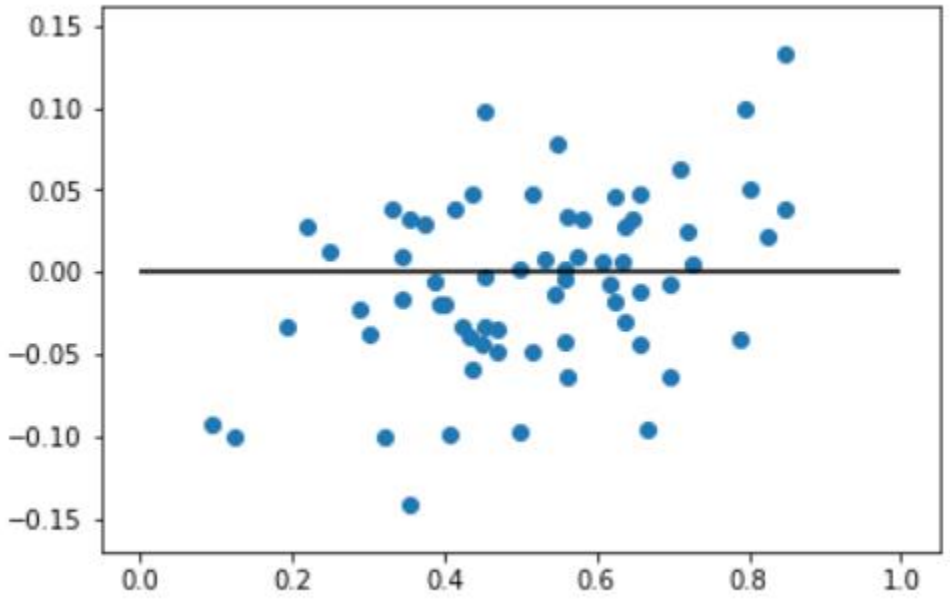
- Gathered regression model coefficients

```
1 print("coefficients: " + str(lrModel.coefficients))
2 print("intercept: " + str(lrModel.intercept))
```

```
coefficients: [-0.0223586565793701 7,0.011511843550962755,-0.0010042426100046833]
intercept: 0.690567775337331
```

Visualizations

- Residual plot for regression on win percentage



- Average team efficiency by conference