FOOTBALL (SOCCER) MATCH RESULT PROJECTIONS

- Nikolai Stambler-Tennant
- Brown University – DSI

- December 7th, 2022
- GitHub - Niktennant

RECAP

The problem?	Importance?	Type?	Data Origin?	Preprocessing/EDA

• Predicting football matches outcomes - specifically in the Premier League.	• Billion $ industry, sports betting, personal interest	• Multi-class Classification (H : 2.0, D : 0.0, A : 0.0)
		• Time Series
			• Multiple sources
			  • Fantasy Premier League API
			  • FIFA 22 & FIFA 23
			  • PL Match Records
				• Preprocess and lag data four times (1, 3, 5 ,7)
				• Correlation between created variables and target

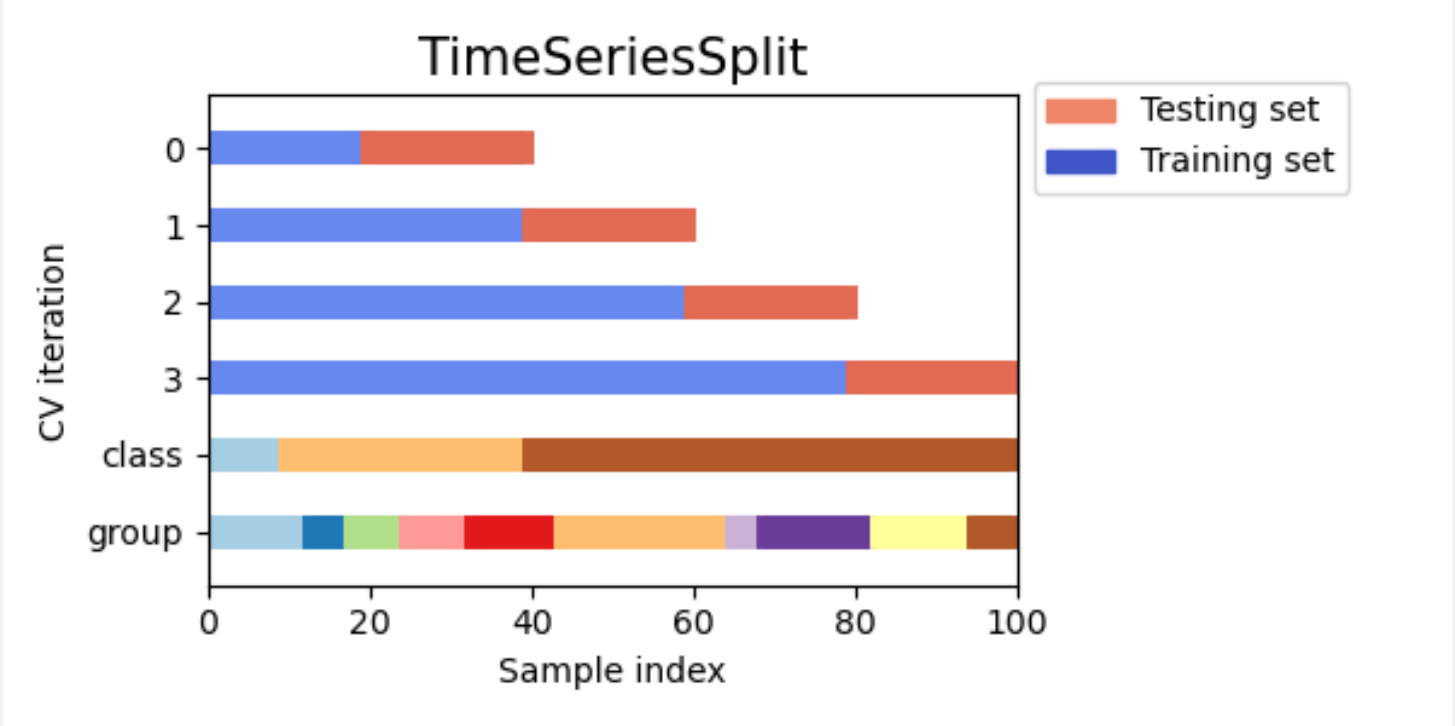
	Referee	B365H	B365D	B365A	HICT	AICT	DifICT	Result
0	M Oliver	4.00	3.40	1.95	7300.0	7900.0	-600.0	2.0
1	A Madley	1.90	3.50	4.00	7900.0	7600.0	300.0	2.0
2	D Coote	3.10	3.10	2.45	7600.0	7600.0	0.0	0.0
3	J Moss	1.25	5.75	13.00	8300.0	7600.0	700.0	2.0
4	M Dean	3.10	3.20	2.37	7500.0	7800.0	-300.0	2.0
...	...	...	...	...	...	...	...	...
452	M Oliver	2.50	3.60	2.62	8323.8	8698.8	-375.0	2.0
453	P Tierney	2.05	3.50	3.60	7795.1	7765.2	29.9	2.0
454	C Kavanagh	1.72	3.80	4.75	8176.2	7793.7	382.5	2.0
455	D Coote	3.60	3.50	2.05	8001.2	8402.0	-400.8	0.0
456	A Taylor	3.30	3.40	2.20	7796.1	8081.8	-285.7	1.0

457 rows × 8 columns

RECAP PT.2

CROSS VALIDATION

- Split with TimeSeriesSplit and used gridsearchcv
- Created multiple algorithms (individual algs for all models)
- L1, L2, CATBOOST, ElasticNet, SVC
- 2 splits due to better performance
- Parameters : all that my computer could handle
- Main focus on Random seed/randoms state due to time series



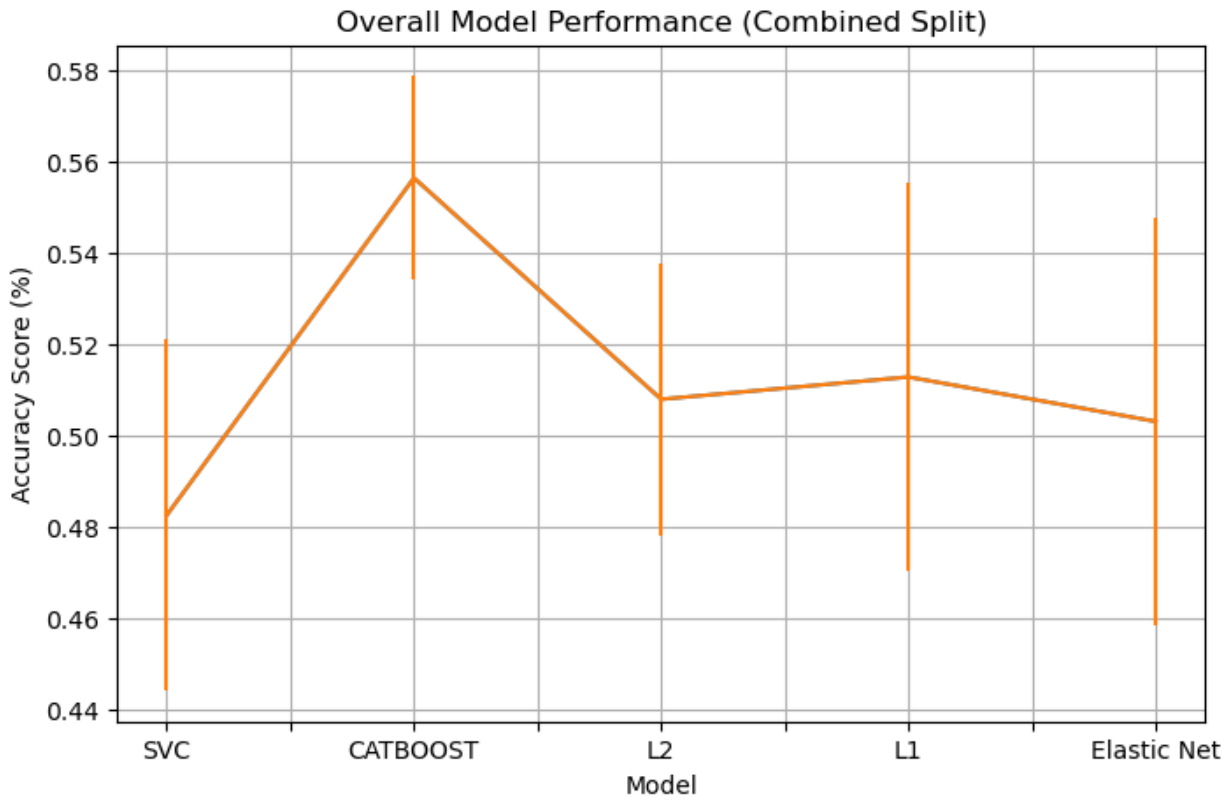
```
tscv = TimeSeriesSplit(n_splits)

for train_index, test_index in tscv.split(X):
    X_train, X_test = X.iloc[train_index], X.iloc[test_index]
    y_train, y_test = y.iloc[train_index], y.iloc[test_index]
    pipe = Pipeline(steps=[('preprocessor', preprocessor), ('ml', model)])
    gsearch = GridSearchCV(pipe, cv=tscv, param_grid= param_grid, scoring = scoring, refit = refit)
    gsearch.fit(X_train, y_train)
```

CROSS VALIDATION PT.2

RESULTS

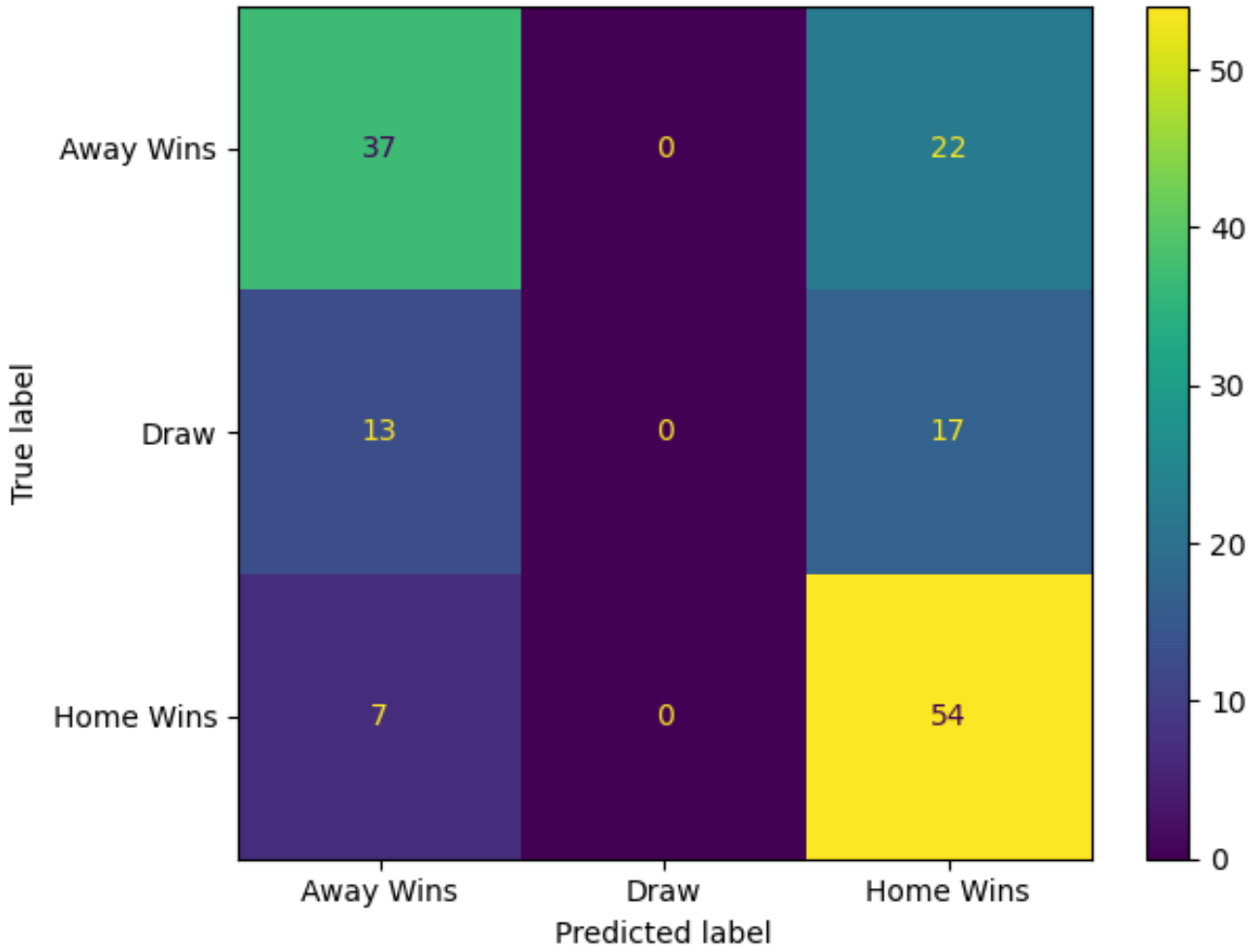
Best Model Accuracy: 0.61
Worst Model Accuracy: 0.43
Baseline: 0.43
Total models: 200
(5 models x 2 splits x 4 lags x 5 rs)

	Mean	STD	SEM
SVC	0.482654	0.038462	0.006081
CATBOOST	0.556552	0.022356	0.003535
L2	0.508061	0.029909	0.004729
L1	0.512931	0.042423	0.006708
Elastic Net	0.503201	0.044560	0.007046



RESULTS PT.2

	Model Type	Lag	Score	RS	Split
0	CATBOOST	Lag 7	0.606667	3	1



OUTLOOK

- More data! Increase from 2 to 20 seasons
- Increase range of hyperparameters
- More features related to scores (average goals, average saves, ect.)