Data Science

Concrete compressive strength prediction

Shao-Ning, Chen

Institute of Data Science, National Cheng Kung University

Jun 8, 2022

Recap

- The dataset comes from UCI Concrete Compressive Strength Data Set [1]
- Number of instances (observations): 1030
- Attribute breakdown: 8 quantitative input variables, and 1 quantitative output variable
- Missing Attribute Values: None

	Cement	Slag	Fly_Ash	Water	Superplastic	Coarse_Aggr	Fine_Aggr	Age	CCStr
0	540.0	1	1	162.0	2.5	1040.0	676.0	2	79.99
1	540.0	1	1	162.0	2.5	1055.0	676.0	2	61.89
2	332.5	3	1	228.0	0.0	932.0	594.0	5	40.27
3	332.5	3	1	228.0	0.0	932.0	594.0	5	41.05
4	198.6	3	1	192.0	0.0	978.4	825.5	5	44.30

Variable Information

Name	Unit	Description	Dtype
Cement	kg/m3	Cement	float
*Blast Furnace Slag	-	Metal oxides and SiO2 mix	cat.
*Fly Ash	-	Coal combustion product	cat.
Water	kg/m3	Water	float
Superplasticizer	kg/m3	Making high-strength concrete	float
Coarse Aggregate	kg/m3	Larger than 4.75mm aggregate	float
Fine Aggregate	kg/m3	Small than 4.75mm aggregate	float
*Age	-	Age	cat.
CCStrength	MPa	Output Variable	float

- Variables contain * has been binned from 1 to 5
- Use label encoding to handle binned categorical value

Methods

- Linear regression and its variant [2]
- Machine learning without tree based methods [2]
- Machine learning with tree based methods
- Training and testing on different testing ratio
- Training and testing on different regression methods
- Compare the result with/without data binning

Methods

- Linear Regression*
- Lasso Regression*
- Ridge Regression*
- ElasticNet Regression
- Decision Tree Regression*
- KNN Regression
- MLP Regression
- SVM Regression
- LightGBM Regression
- XGBoost Regression
- CatBoost Regression
- Random Forest Regression*

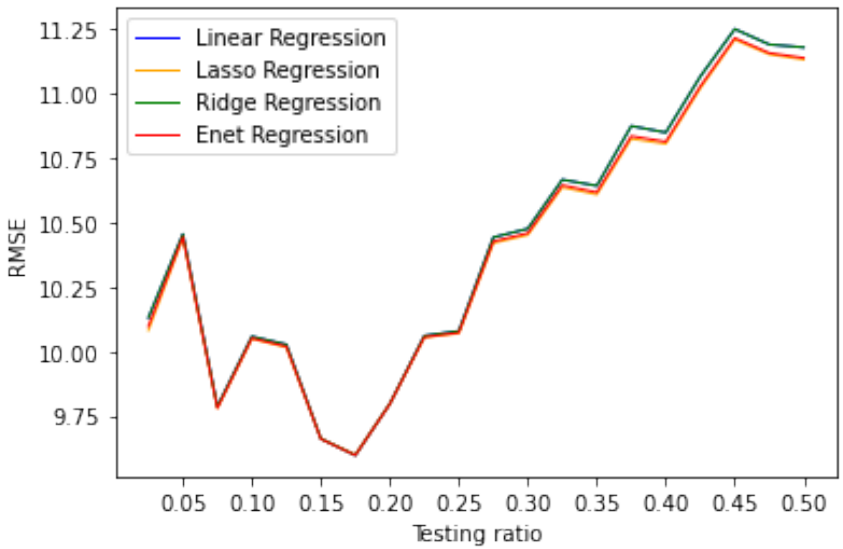
Linear Regression and Variant

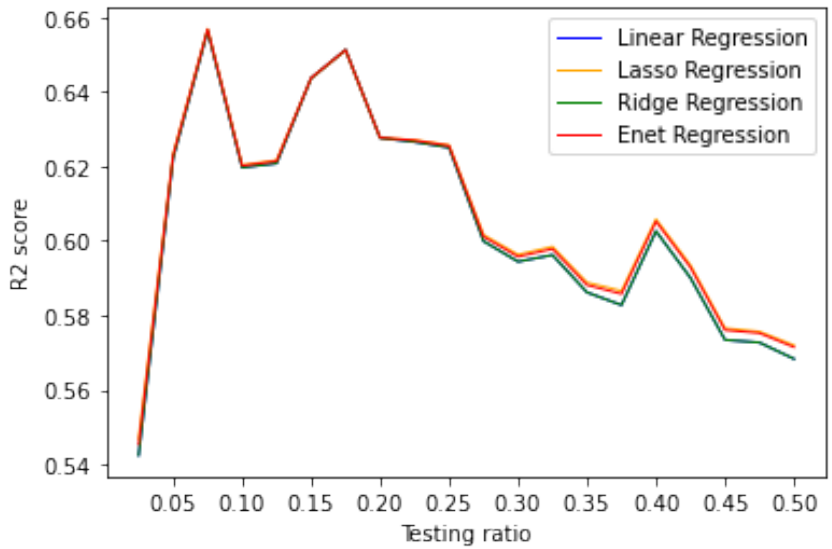
Methods

- Linear Regression*
- Lasso Regression*
- Ridge Regression*
- ElasticNet Regression

Linear Regression and Variant

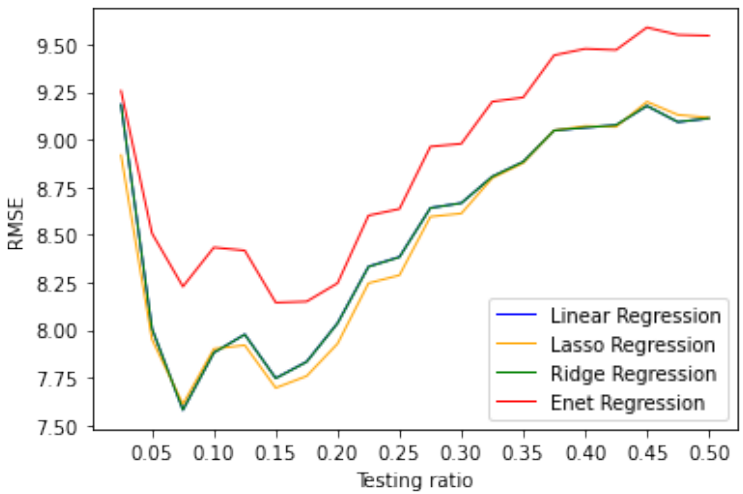
- The test size set between 0.025 to 0.5, and compare the performance between different testing ratio
- The following plots are the performance without data binning

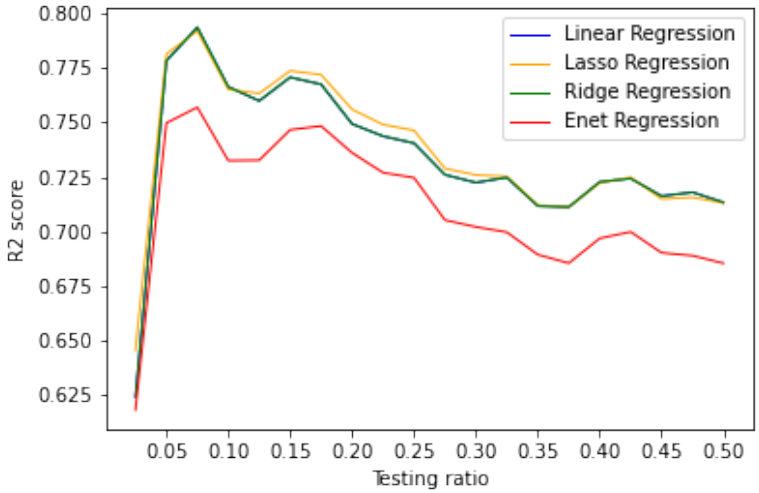


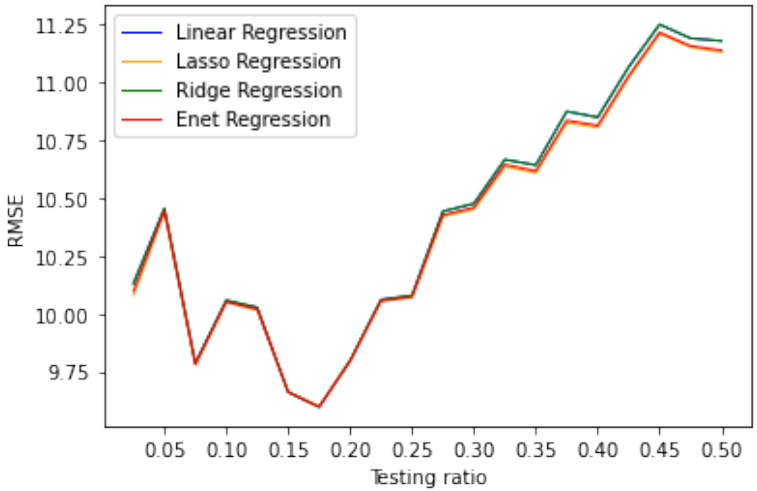


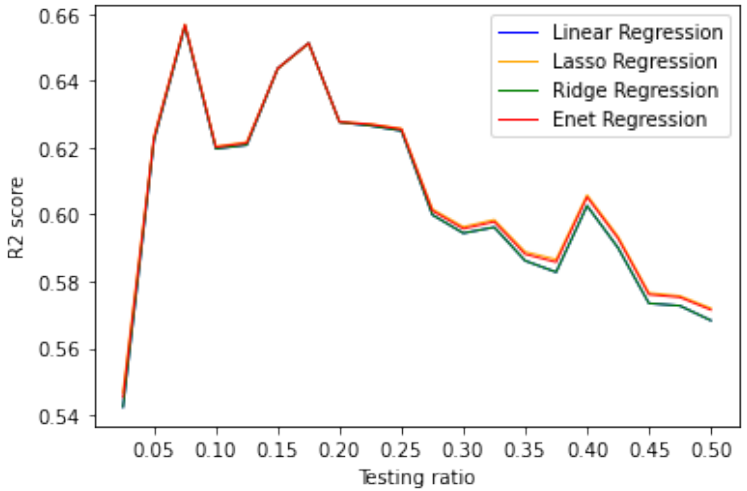
Linear Regression and Variant

- Performance with(upper)/without(lower) data binning



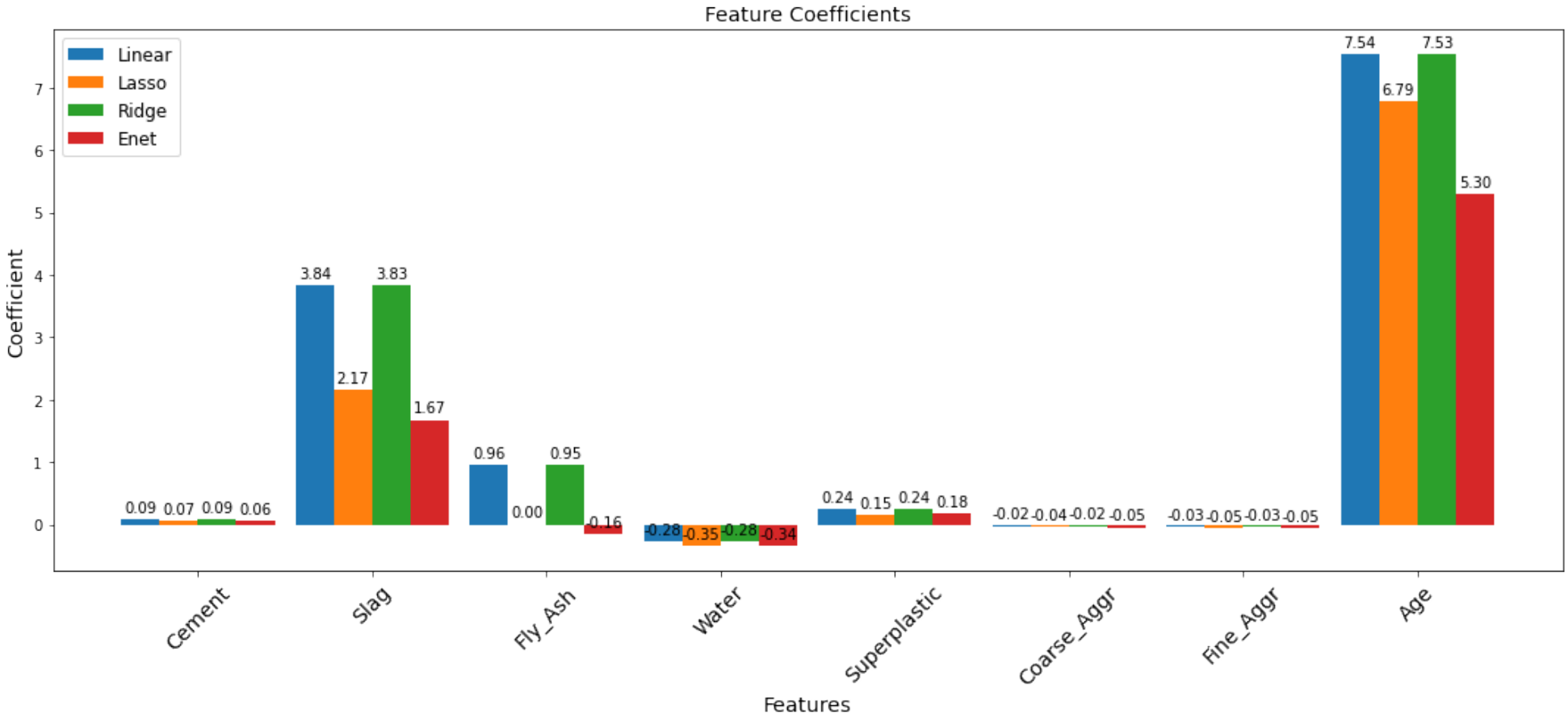






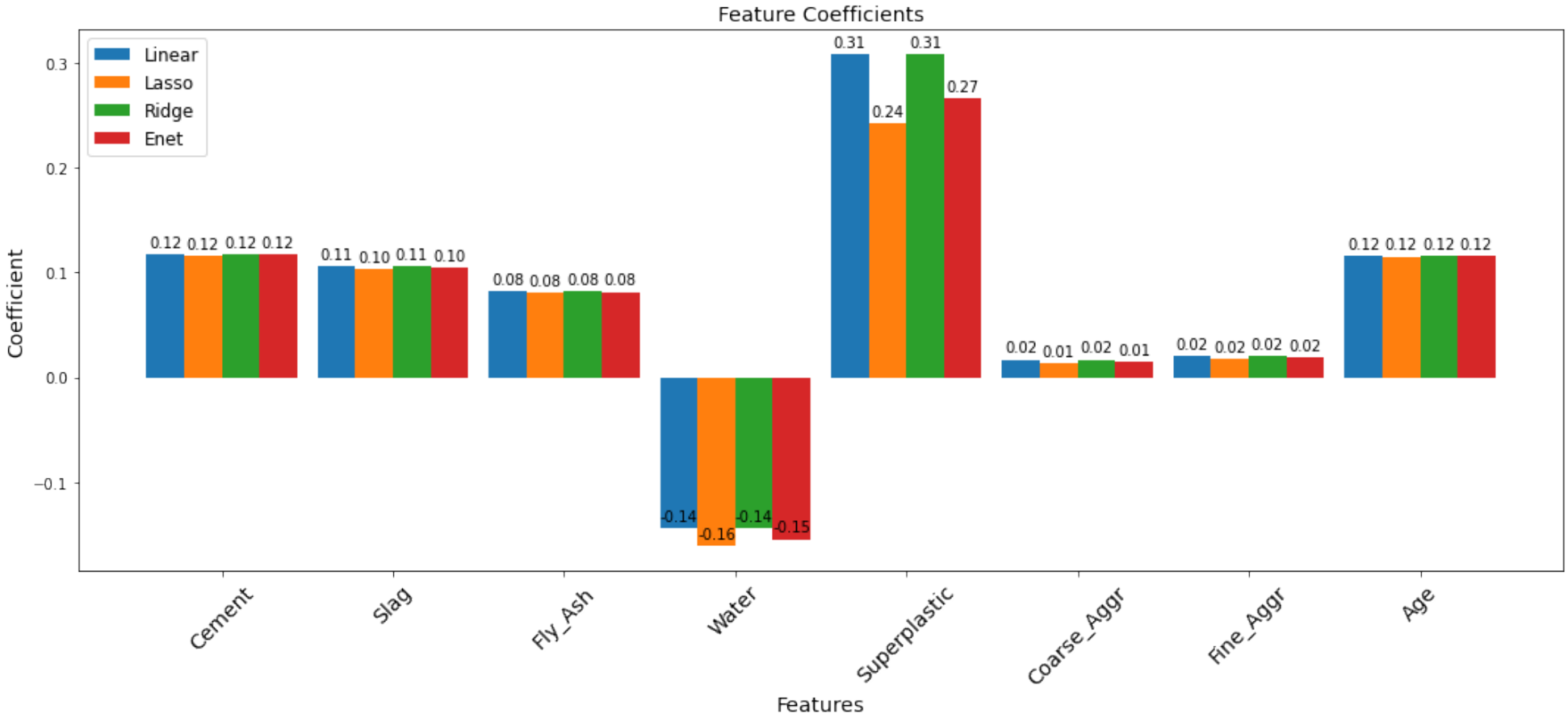
Linear Regression and Variant

- Coefficients with binning dataset (testing ratio=0.15)



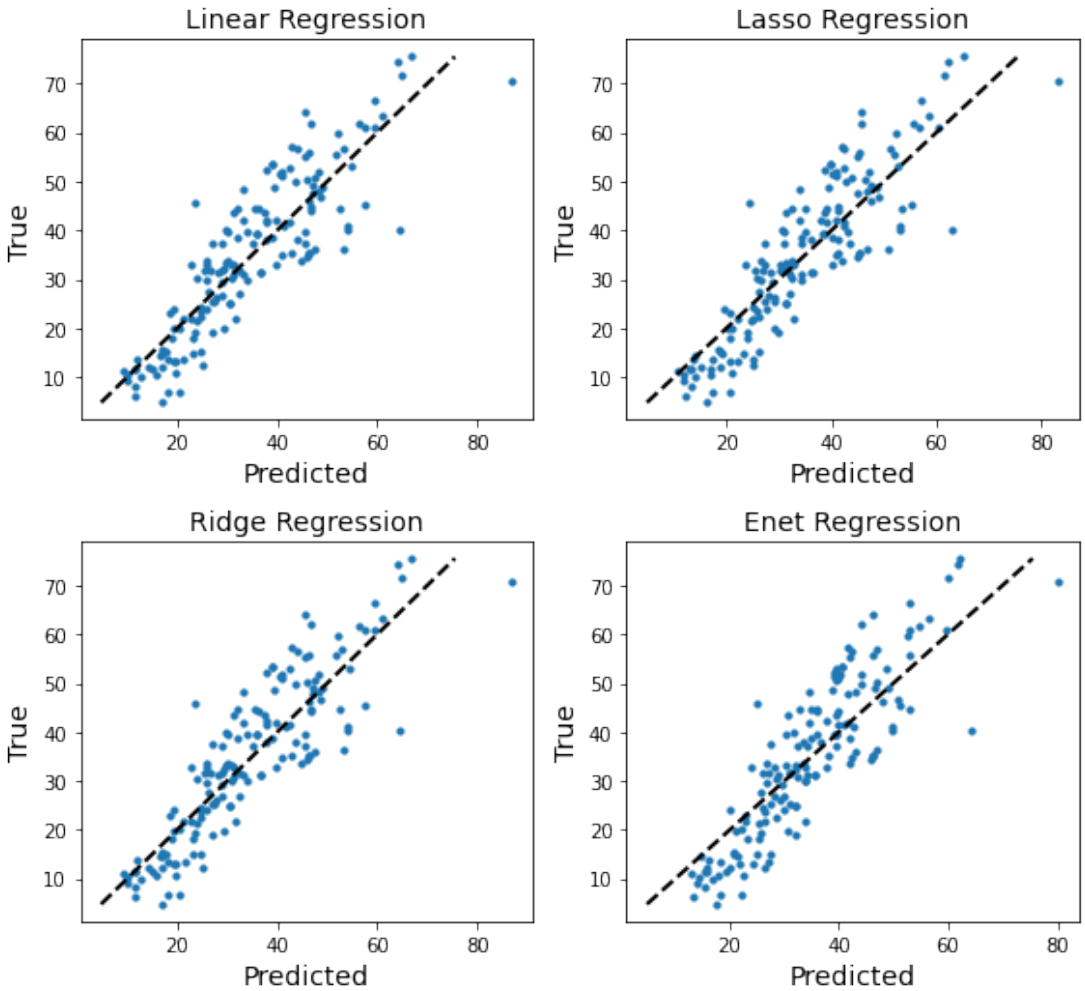
Linear Regression and Variant

- Coefficients without binning dataset (testing ratio=0.15)



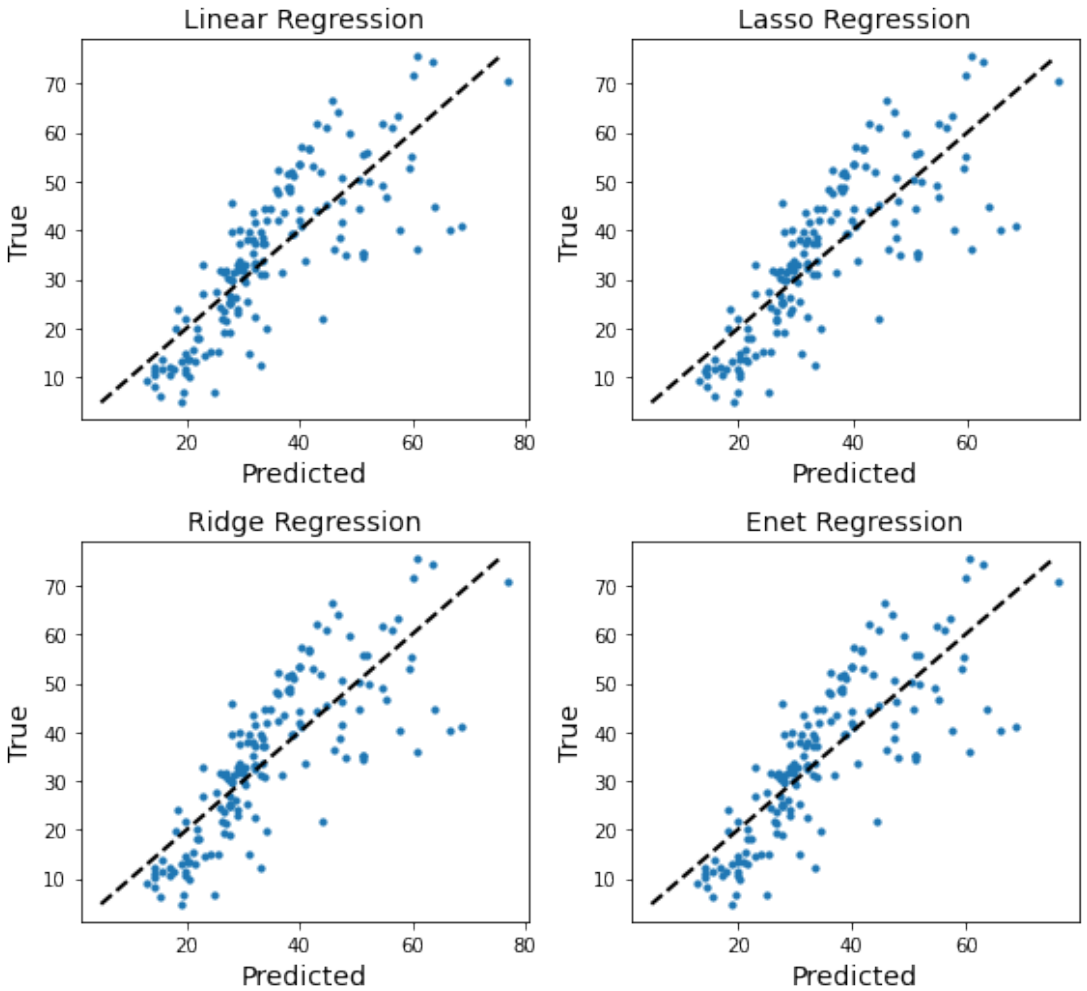
Linear Regression and Variant

- Predicted value vs True value (with data binning, test=0.15)



Linear Regression and Variant

- Predicted value vs True value (without binning, test=0.15)



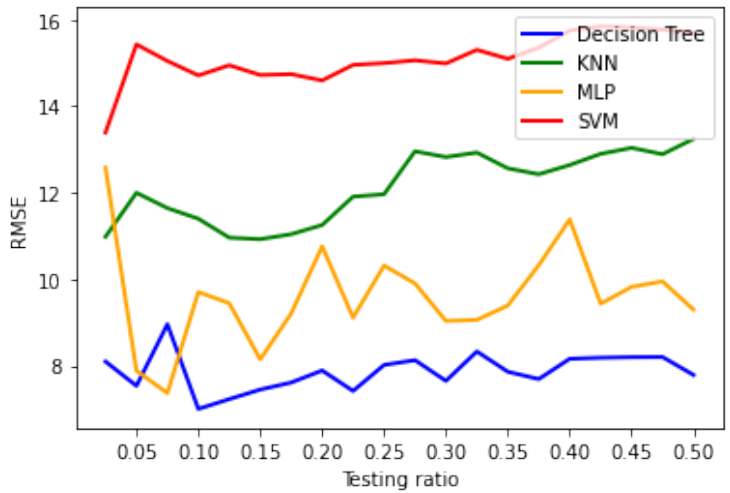
Machine Learning Methods

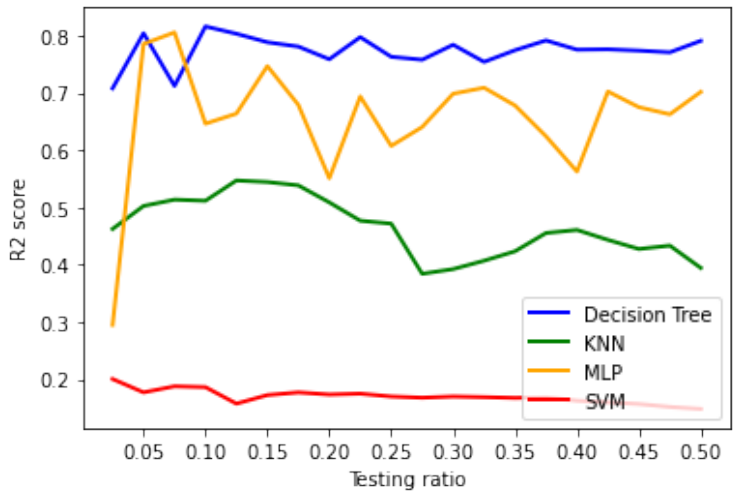
Methods

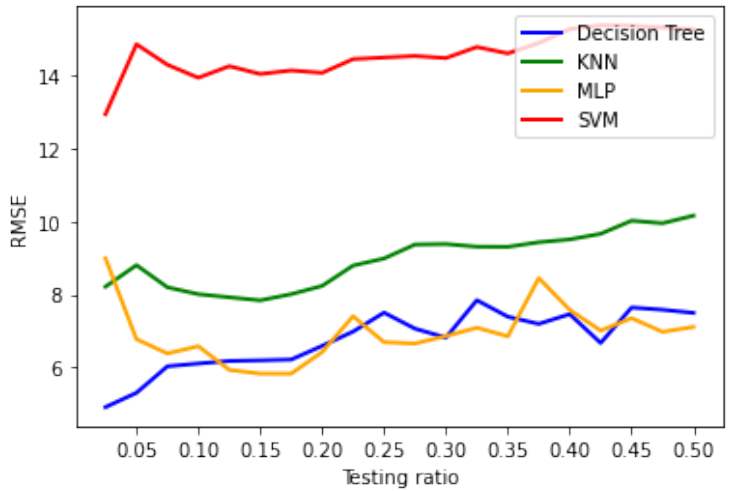
- Decision Trees Regression*
- KNN Regression
- MLP Regression
- SVM Regression

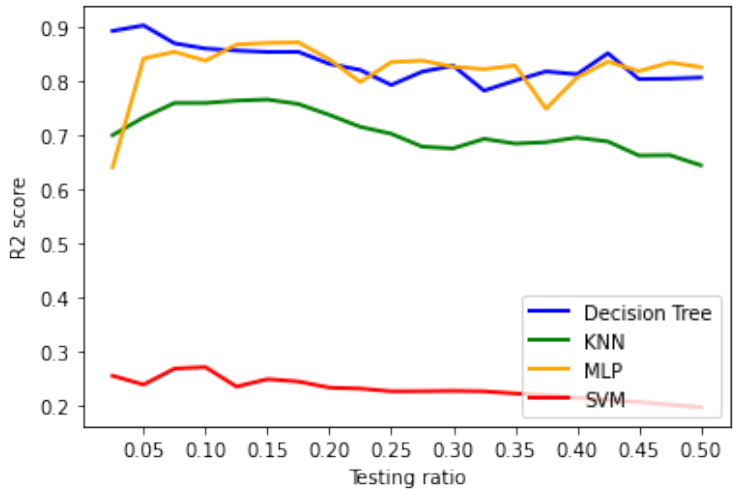
Machine Learning Methods

- Performance with(upper)/without(lower) data binning



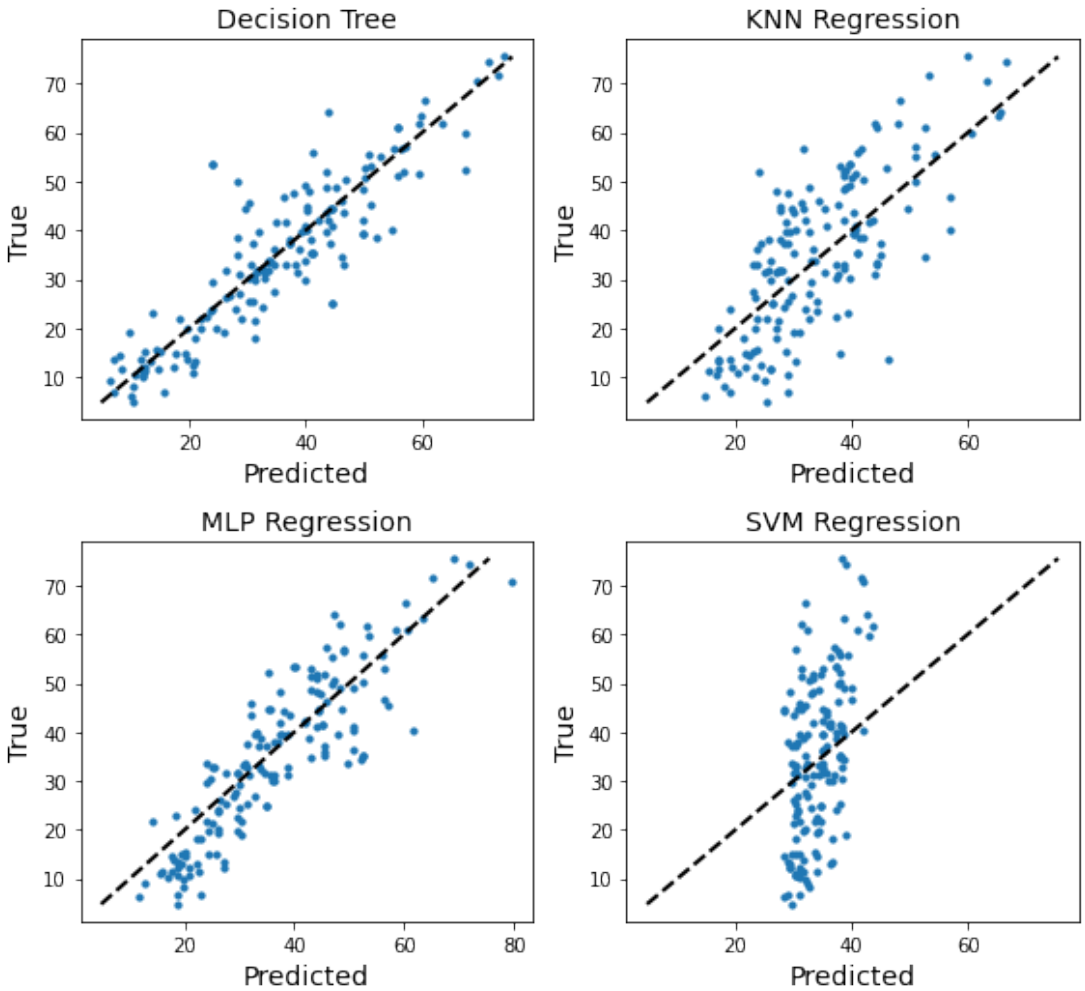






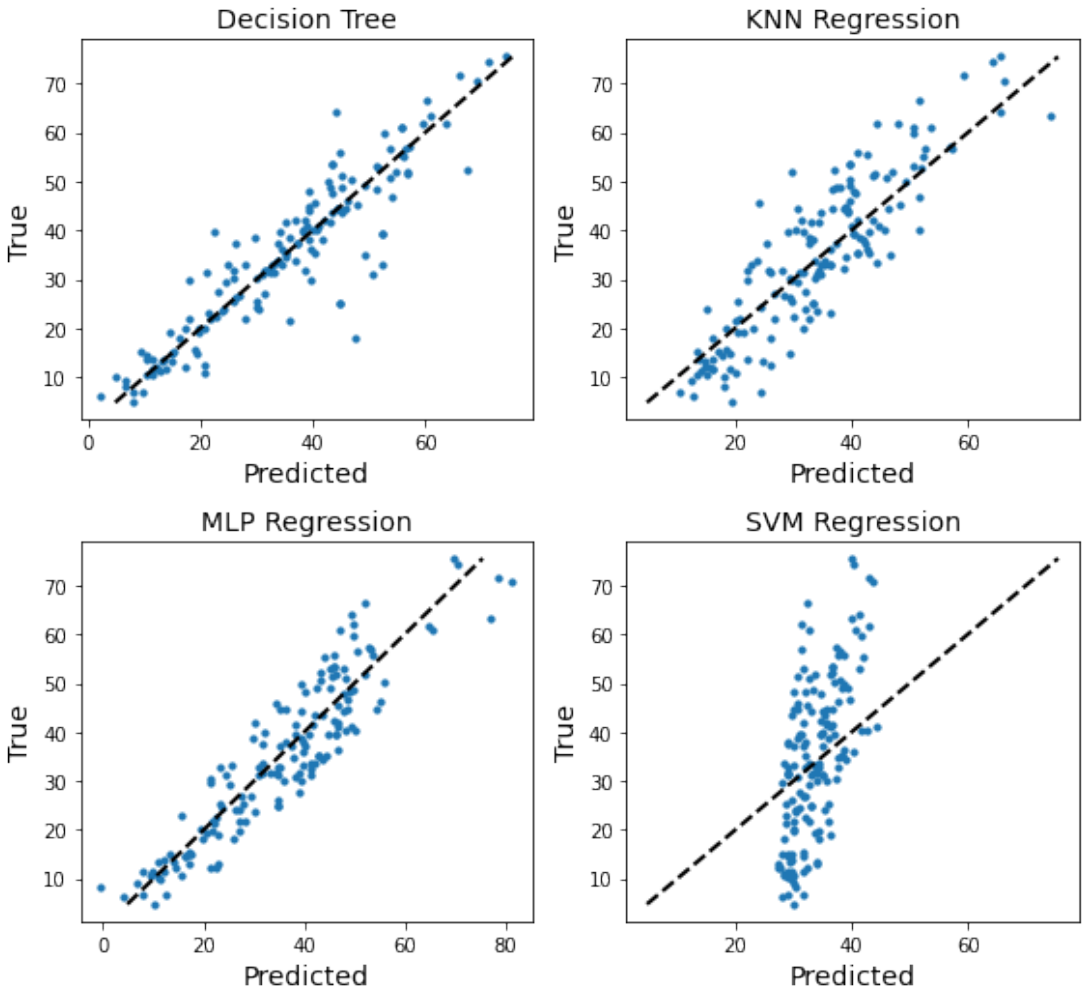
Machine Learning Methods

- Predicted value vs True value (with data binning, test=0.15)



Machine Learning Methods

- Predicted value vs True value (without binning, test=0.15)



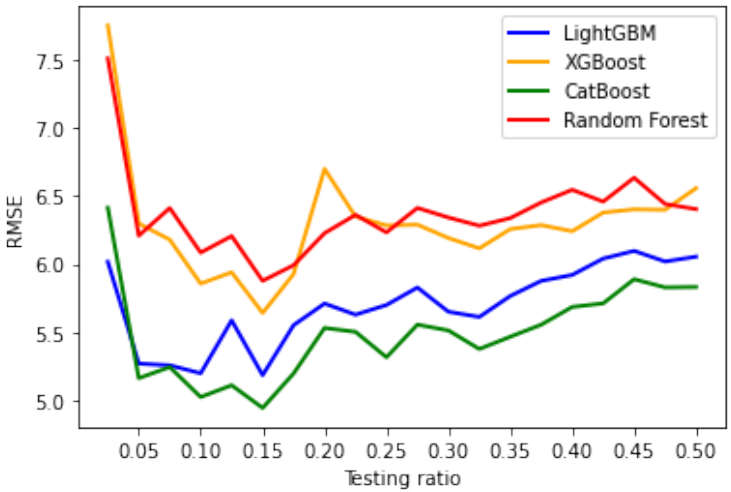
Tree based Methods

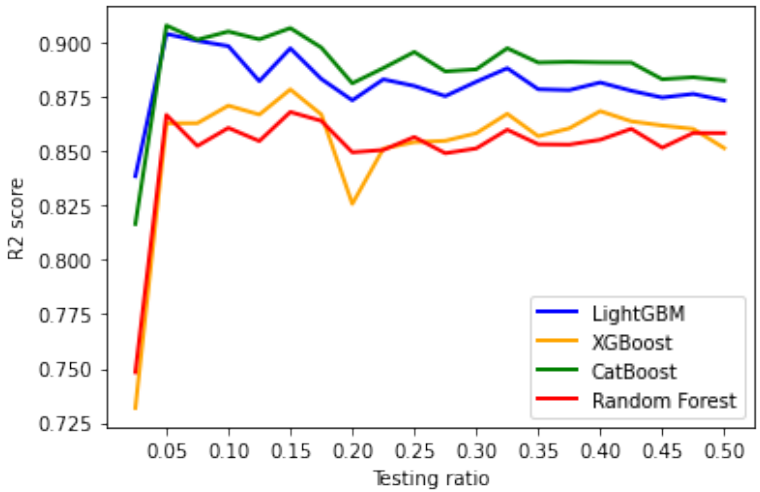
Methods

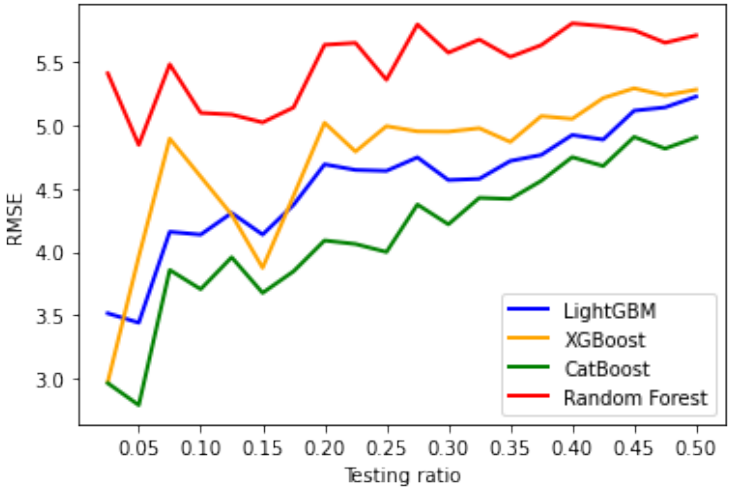
- LightGBM Regression
- XGBoost Regression
- CatBoost Regression
- Random Forests Regression*

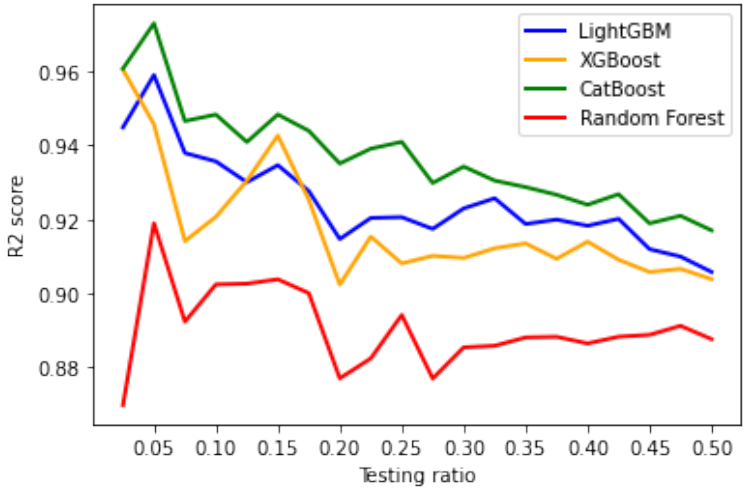
Tree based Methods

- Performance with(upper)/without(lower) data binning



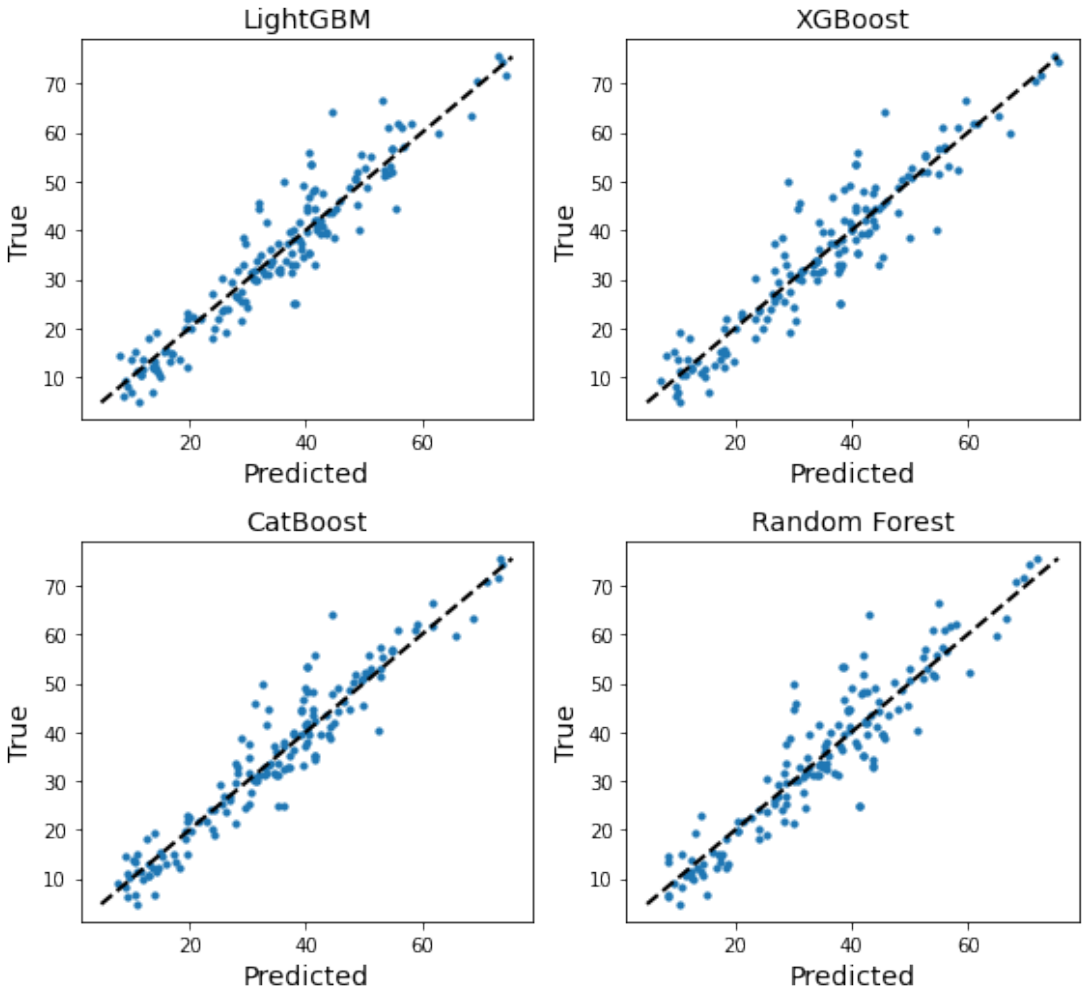






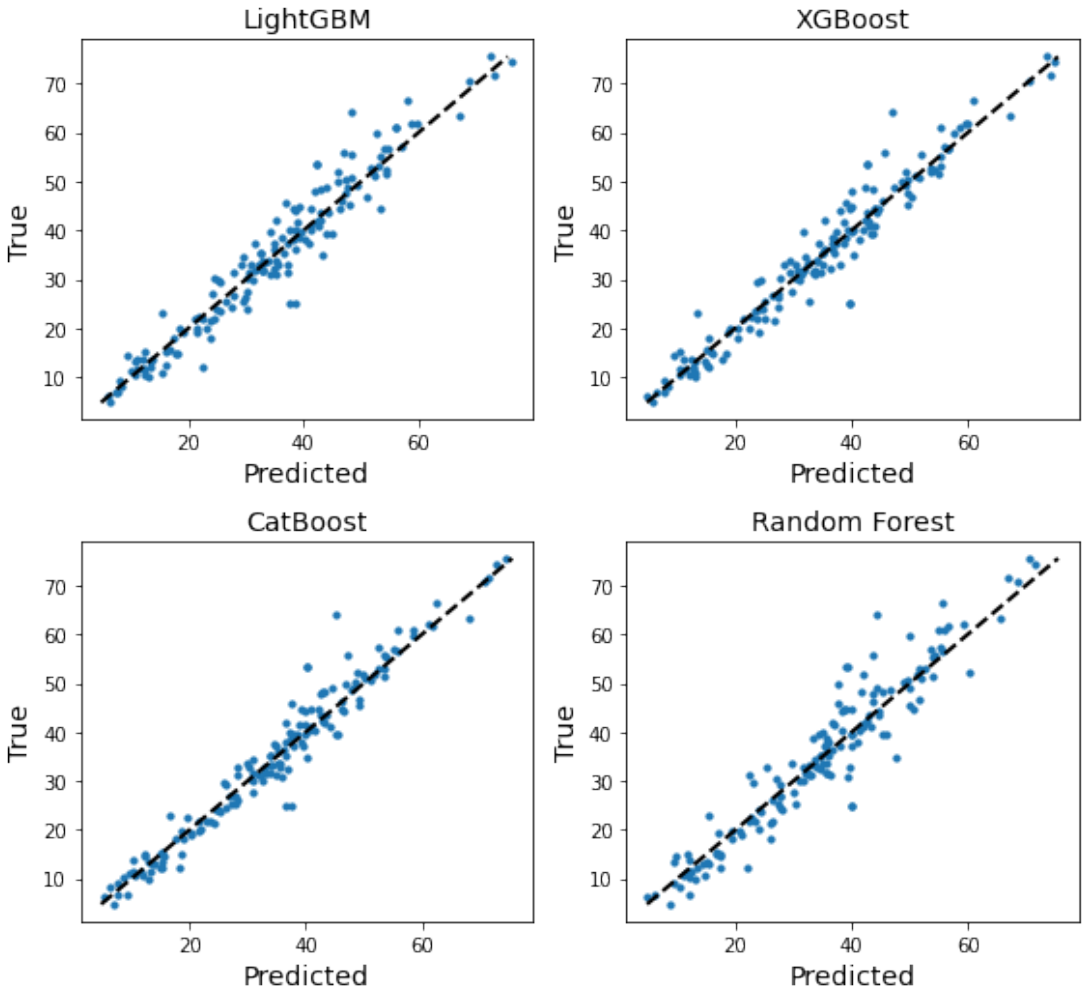
Tree based Methods

- Predicted value vs True value (with data binning, test=0.15)



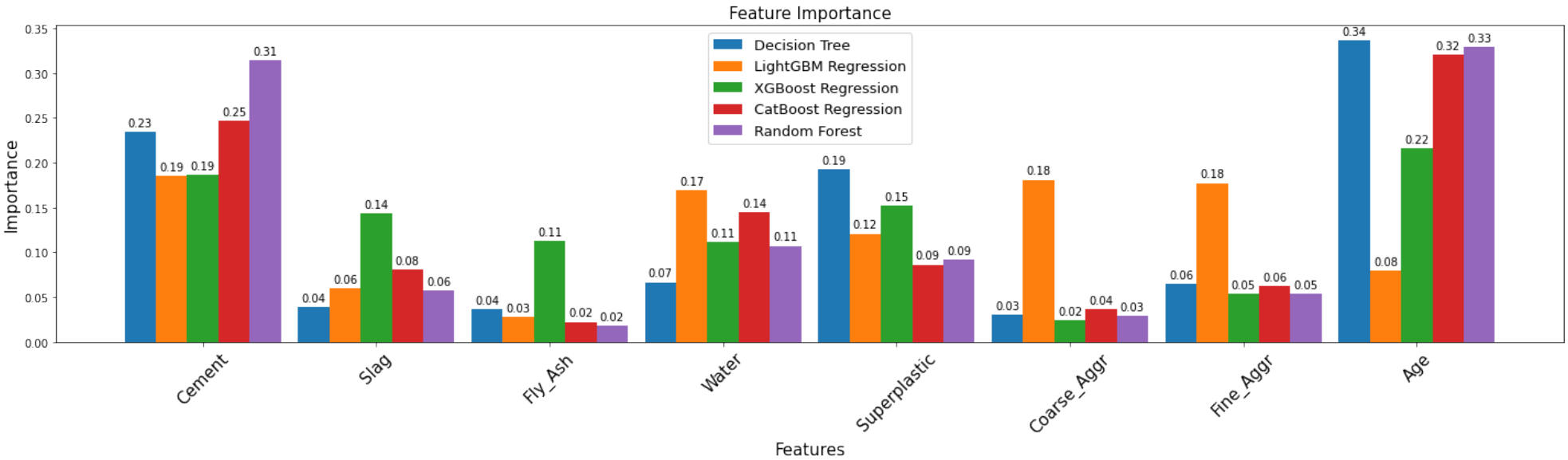
Tree based Methods

- Predicted value vs True value (without binning, test=0.15)



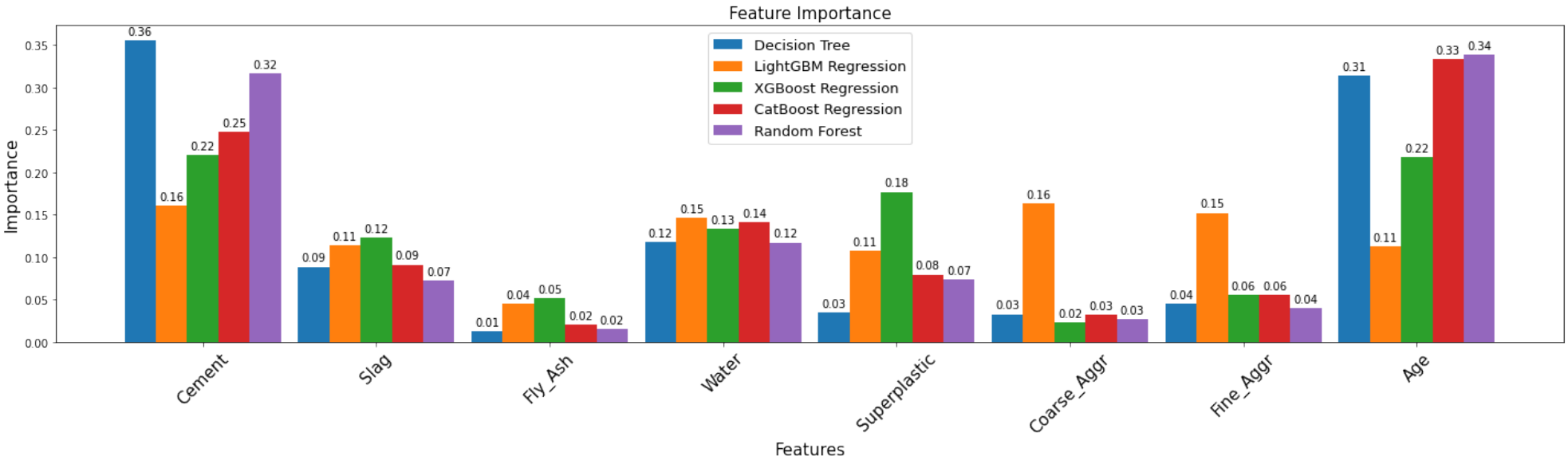
Tree based Methods

- Feature importance of tree based methods (with data binning, test=0.15)



Tree based Methods

- Feature importance of tree based methods (without data binning, test=0.15)



Conclusion

- The lower RMSE occur between testing ratio of 0.05 and 0.2
- In linear regression and its variant, data binning can improve performance
- In machine learning or Tree based methods, data binning can not improve performance
- In Tree baesd methods, as the testing ratio growth, the RMSE without data binning increase rapidly than the dataset with data binning
- Performance : Tree based methods $>$ Linear Regression and variant $\geq$ Machine Learning without tree based methods
- Best model : CatBoost

Reference

[1] I.-C. Yeh. "Modeling of strength of high-performance concrete using artificial neural networks". In: *Cement and Concrete Research* 28.12 (1998), pp. 1797–1808. ISSN: 0008-8846. DOI: https://doi.org/10.1016/S0008-8846(98)00165-3. URL: https://www.sciencedirect.com/science/article/pii/S0008884698001653.

[2] Ahsanul Kabir, Monjurul Hasan, and Md Khasro Miah. "Strength prediction model for concrete". In: *International Journal on Civil and Environmental Engineering* 2.1 (2013), p. 14.

Thank you!