HOUSE PRICES BY REGRESSION

Master M1 – 2023/2024

Hands-On Machine Learning: Scikit-Learn

Pablo Mollá, Chen Junjie and Pavlo Poliuha

AGENDA

1. INTRODUCTION
2. DATA ANALYSIS
3. DATA PREPROCESSING
4. DATASET SPLIT
5. MODELS + IMPLEMENTATION
6. METRIC'S EVALUATION
7. RESULTS

AMES HOUSING DATASET: AMERICAN STATISTICAL ASSOCIATION

	Id	MSSubClass	MSZoning	LotFrontage	LotArea	Street	Alley	LotShape
0	1	60	RL	65.0	8450	Pave	NaN	Reg
1	2	20	RL	80.0	9600	Pave	NaN	Reg
2	3	60	RL	68.0	11250	Pave	NaN	IR1
3	4	70	RL	60.0	9550	Pave	NaN	IR1
4	5	60	RL	84.0	14260	Pave	NaN	IR1
5	6	50	RL	85.0	14115	Pave	NaN	IR1
6	7	20	RL	75.0	10084	Pave	NaN	Reg
7	8	60	RL	NaN	10382	Pave	NaN	IR1
8	9	50	RM	51.0	6120	Pave	NaN	Reg
9	10	190	RL	50.0	7420	Pave	NaN	Reg

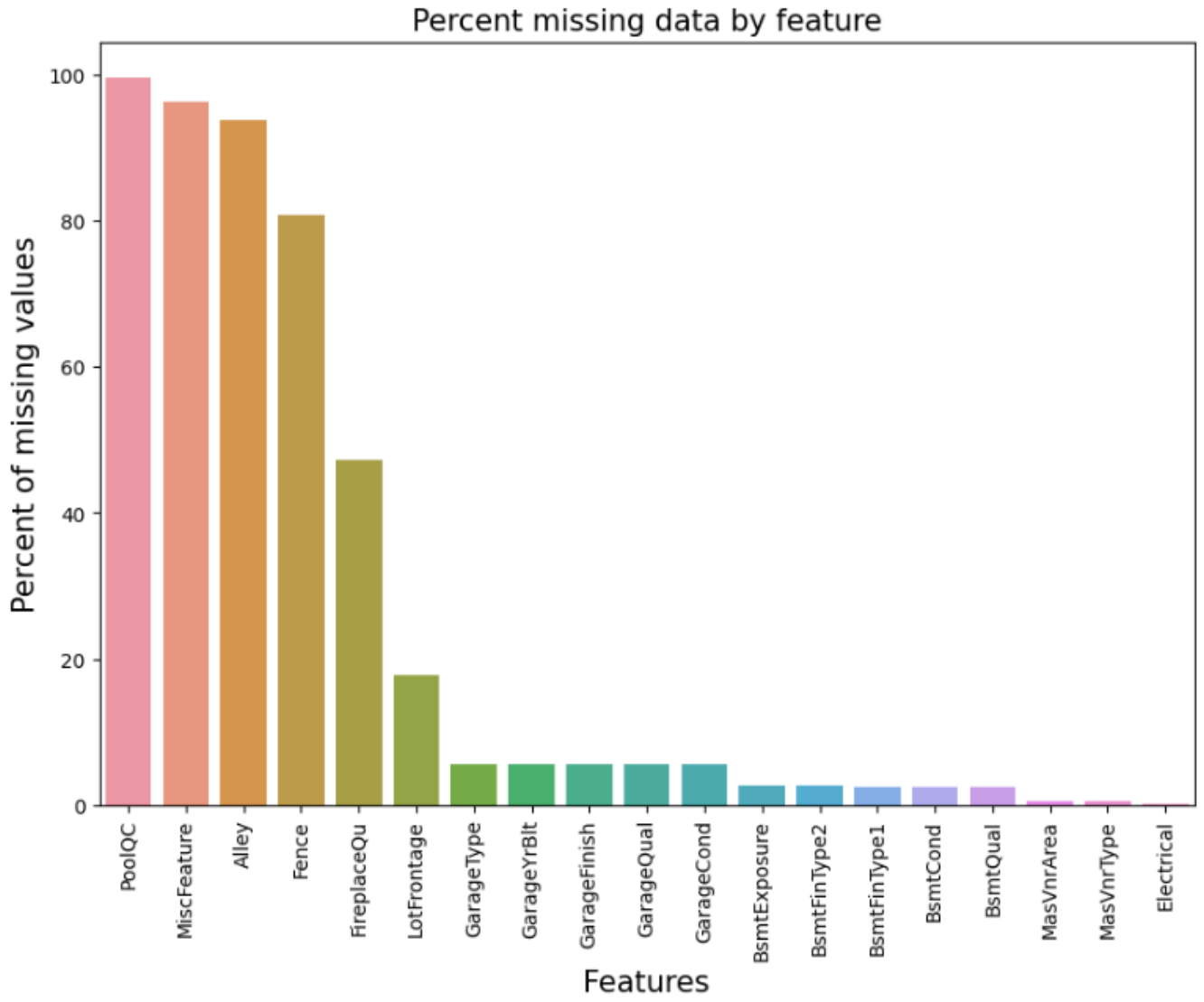
10 rows × 81 columns

MACHINE LEARNING ALGORITHMS:

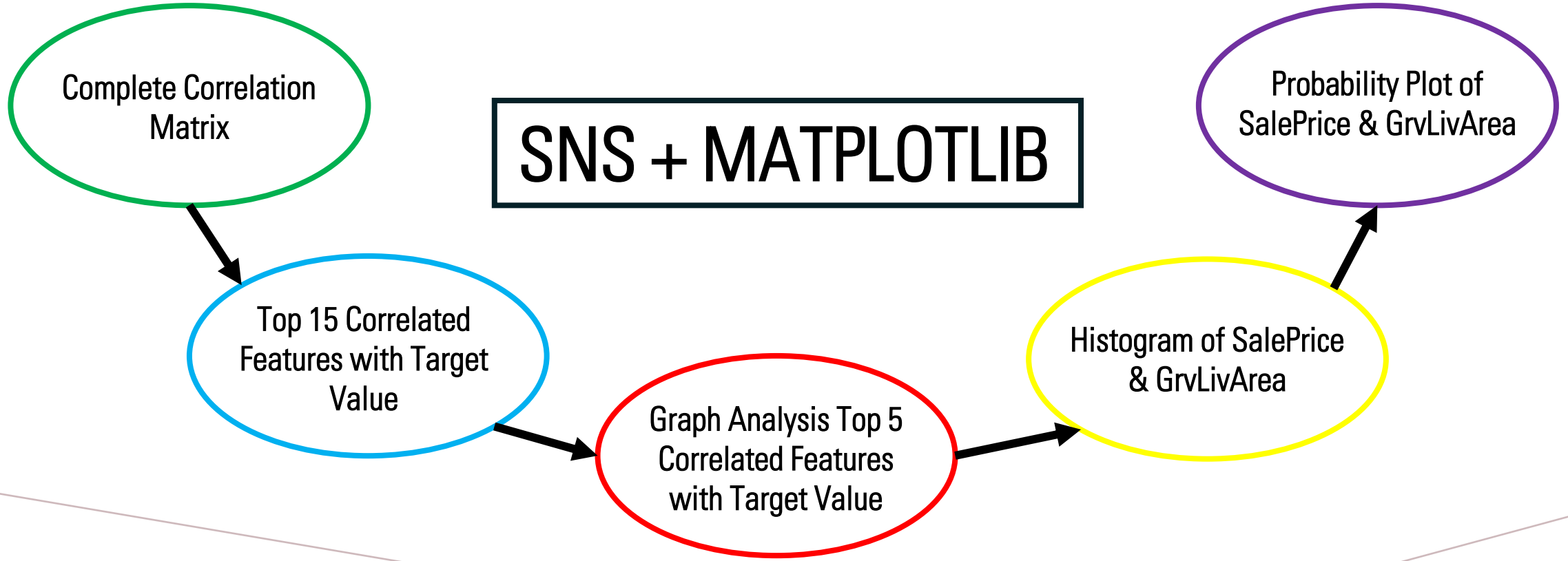
1. LINEAR REGRESSION
2. POLYNOMIAL REGRESSION
3. RIDGE REGRESSION
4. LASSO REGRESSION
5. RANDOM FOREST
6. SUPPORT VECTOR MACHINE (SVM)

DATA CLEANING

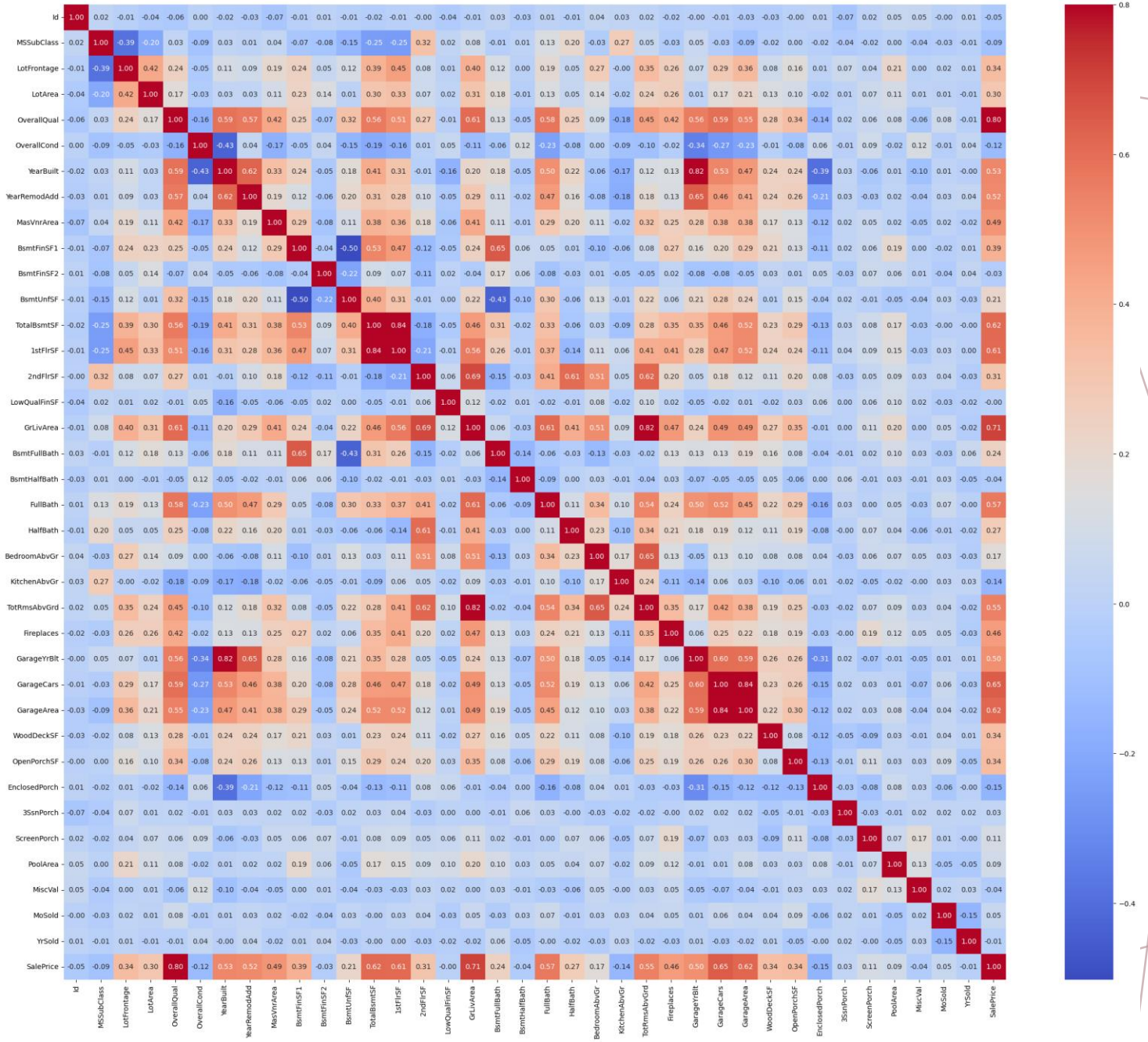
- NUMERICAL FEATURE SELECTION
- DROPPING NAN VALUES



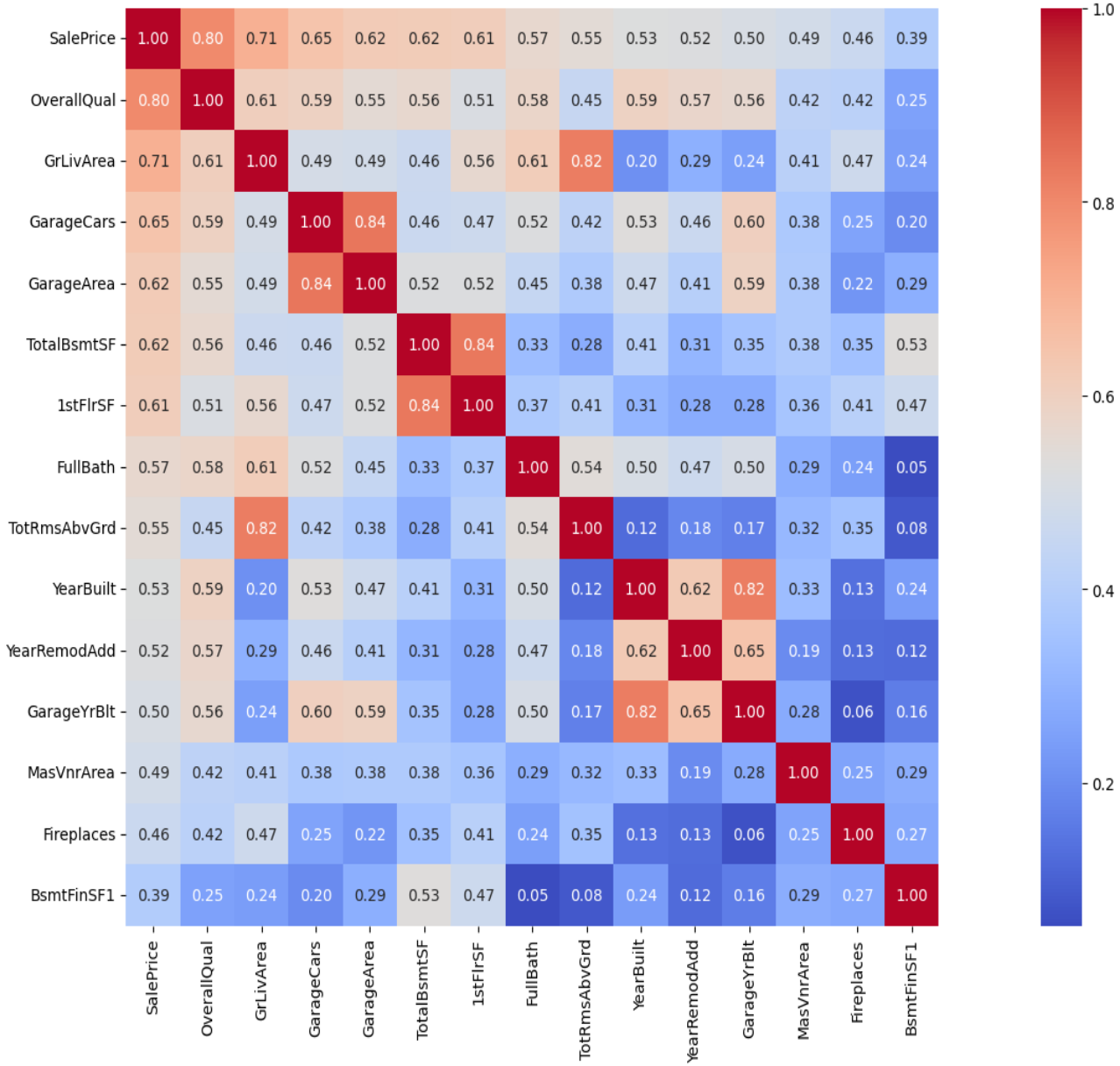
DATA ANALYSIS



CORRELATION MATRIX OF THE NUMERICAL VALUES



MOST CORRELATED FEATURES WITH "SALEPRICE"

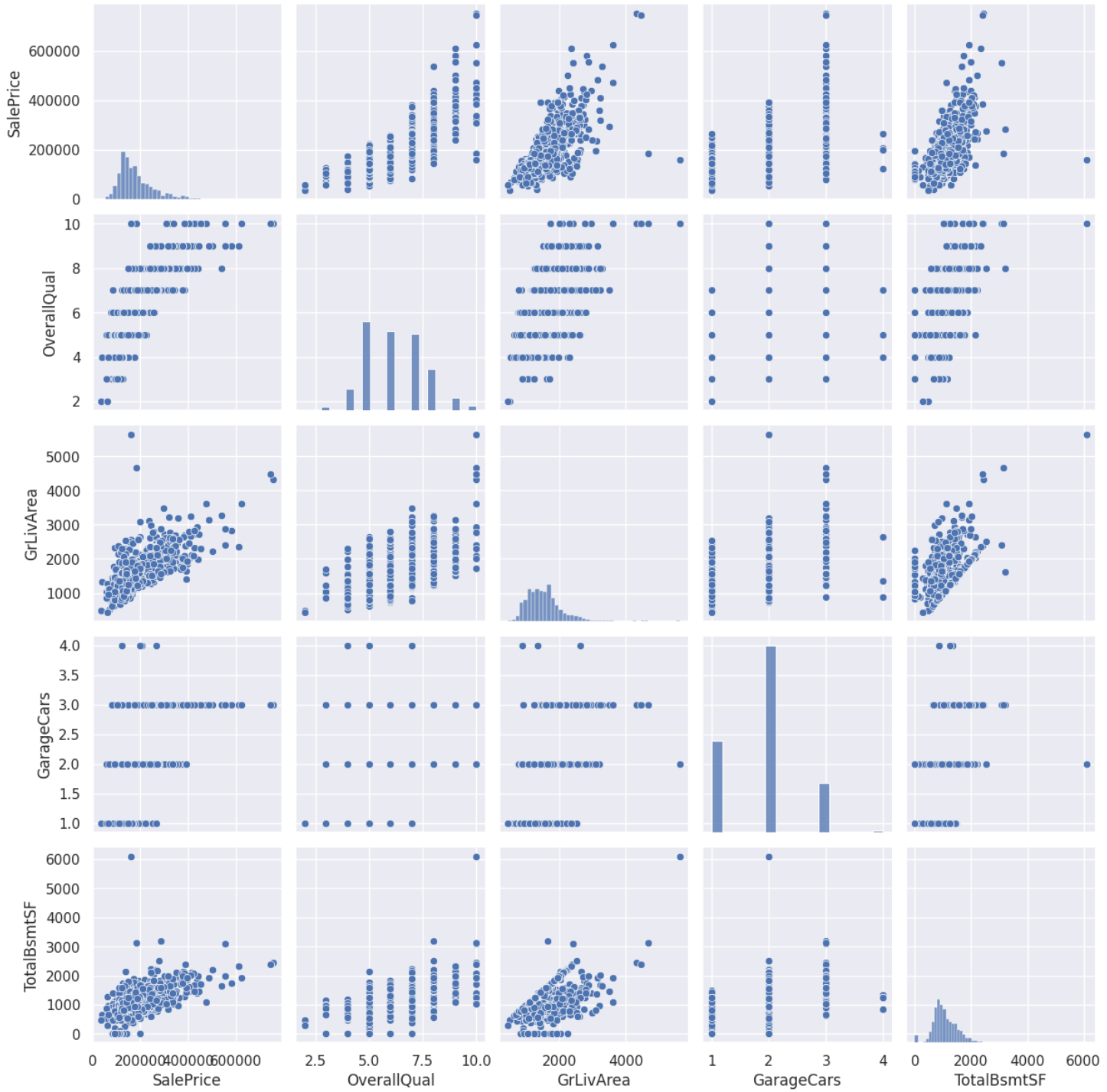


TOP 15 CORRELATED FEATURES:

- Overall Quall
- GrLivArea
- TotRmsAbvGrd
- GarageCars
- GarageArea
- FullBath
- 1stFlrSF
- TotalBsmtSF
- YearBuilt
- MasVnrArea
- Fireplaces
- YearRemodAdd
- LotArea
- BsmtFinSF1

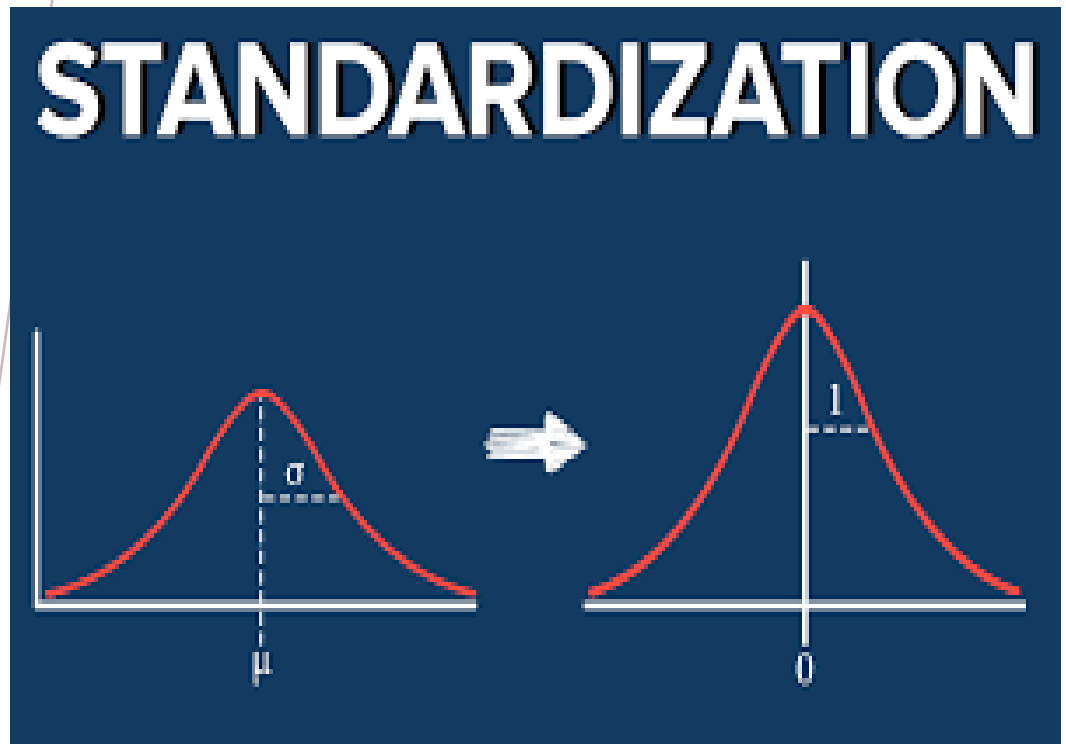
GRAPH ANALYSIS:

TOP 5 MOST CORRELATED FEATURES WITH SALEPRICE



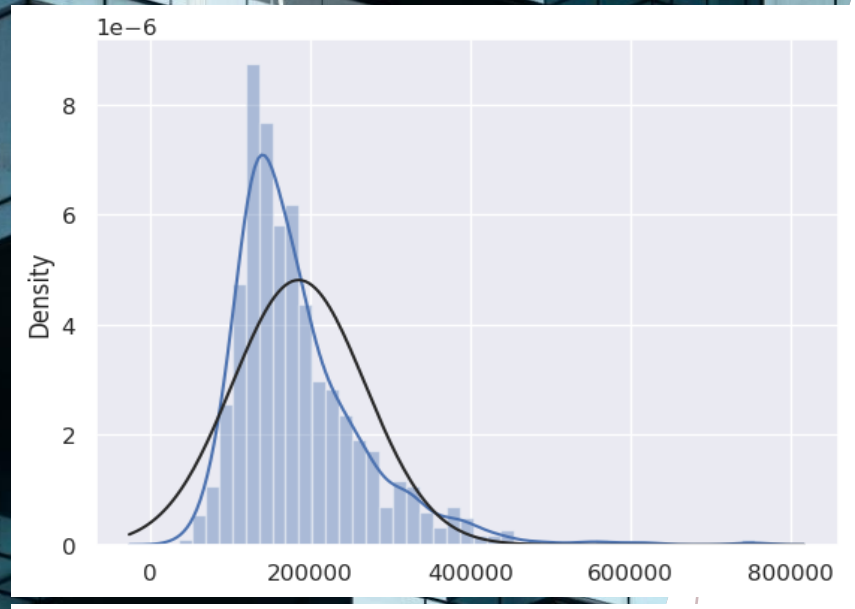
DATA PREPROCESSING

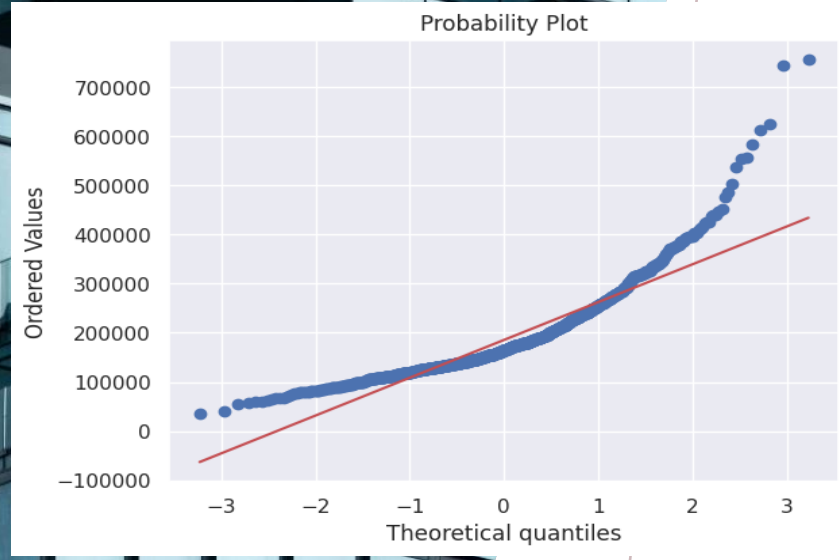
- STANDARDIZATION



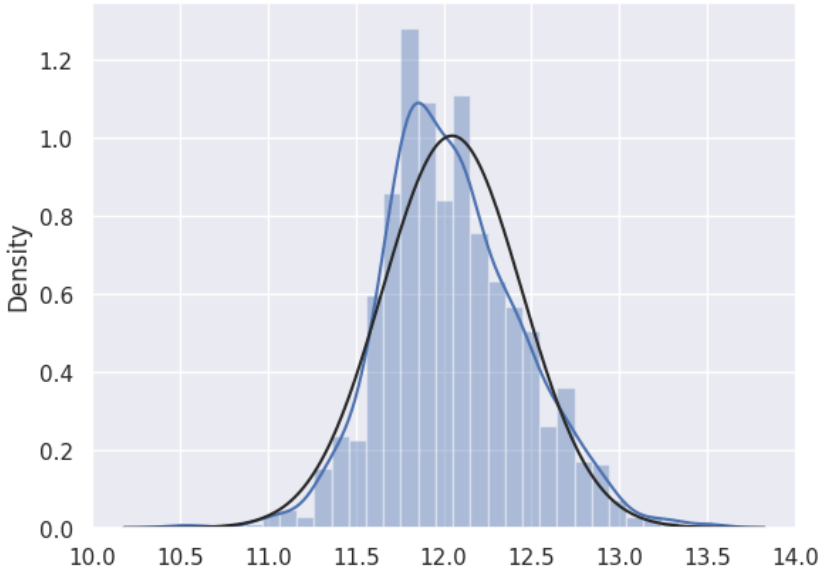
- LOGARITHM APPLICATION

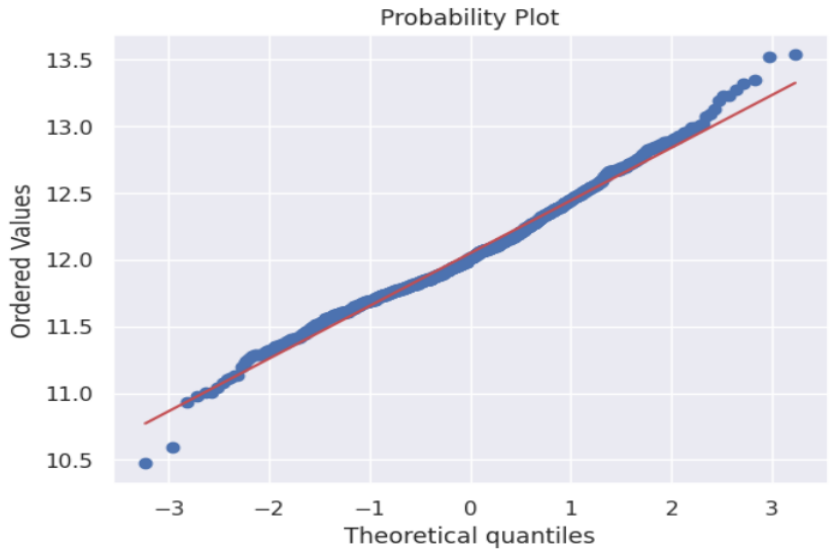
LOGARITHM APPLICATION



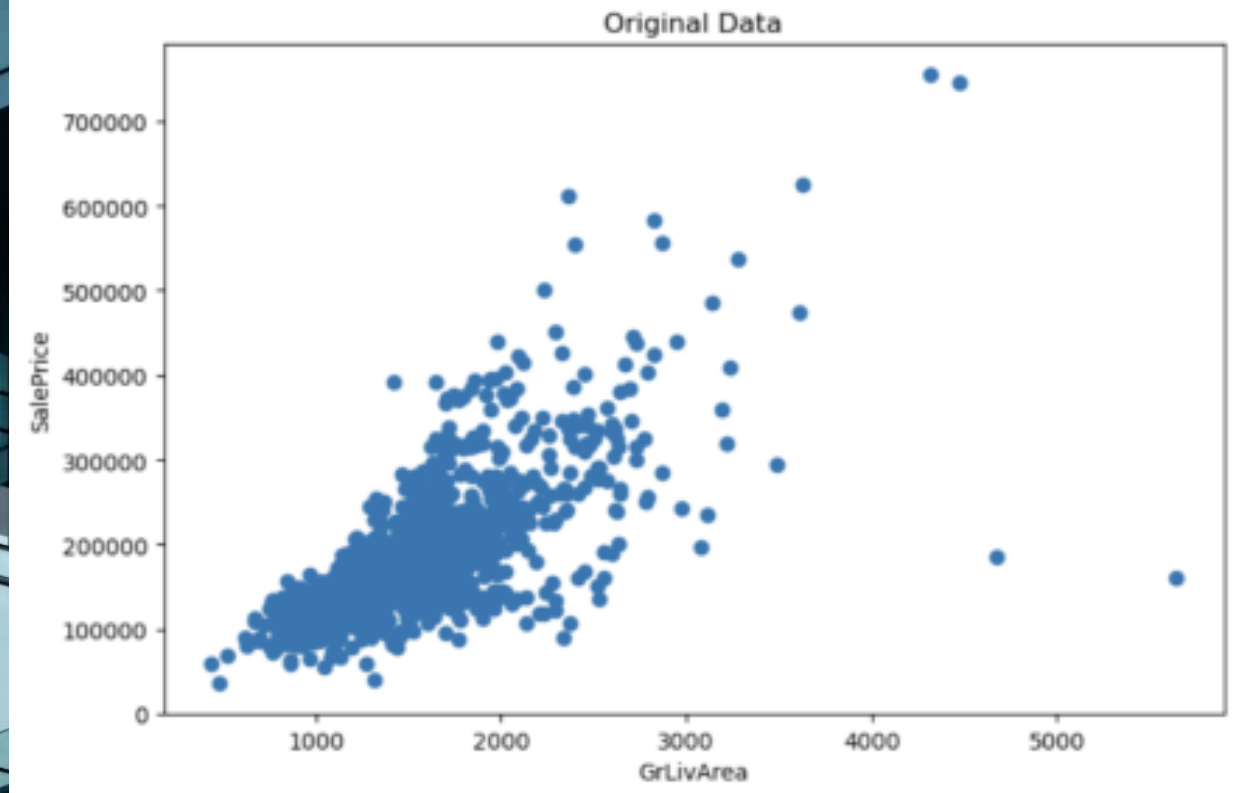


Before / After



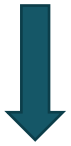


LOGARITHM APPLICATION



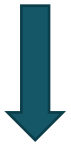
TRAINING / TEST SPLIT

Train Dataset



896 instances

Test Dataset



225 instances

Cross Validation
K=5 Folds

MODELS

LINEAR REGRESSION

POLYNOMIAL REGRESSION

RIDGE REGRESSION

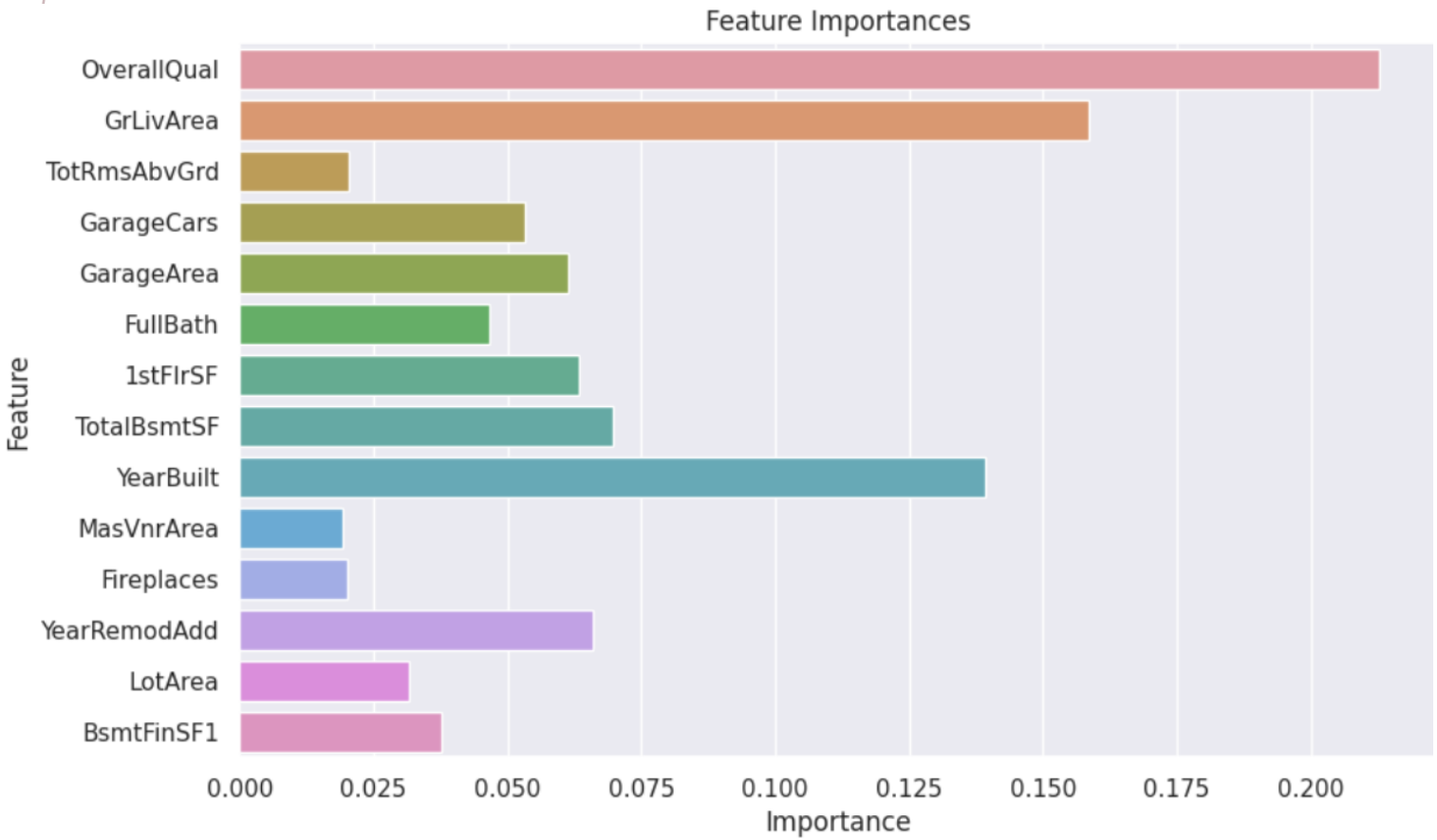
LASSO REGRESSION

RANDOM FOREST

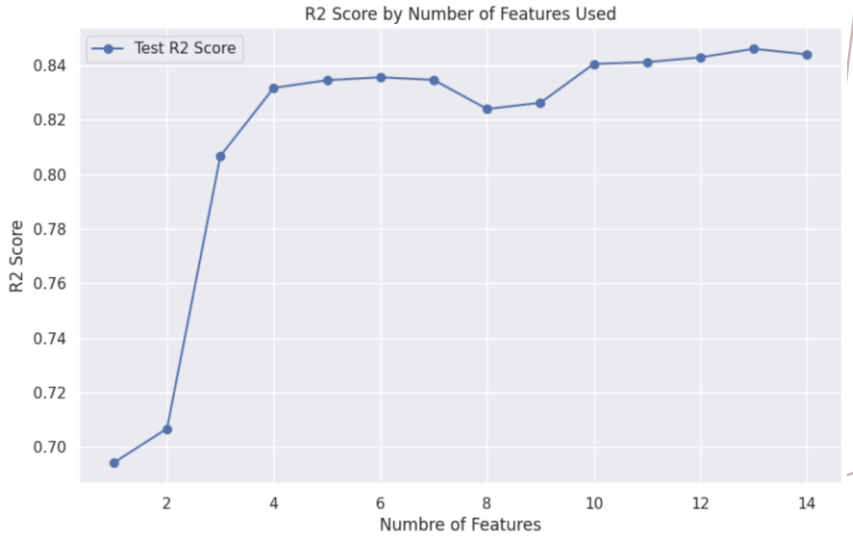
SUPPORT VECTOR MACHINE (SVM)

SELECTION OF THE BEST PARAMETERS USING GRIDSEARCHCV

RANDOM FOREST FEATURE SELECTION



By iteratively screening different features, I found that when removing the feature with the least importance, the performance of the model increased through R2 analysis.



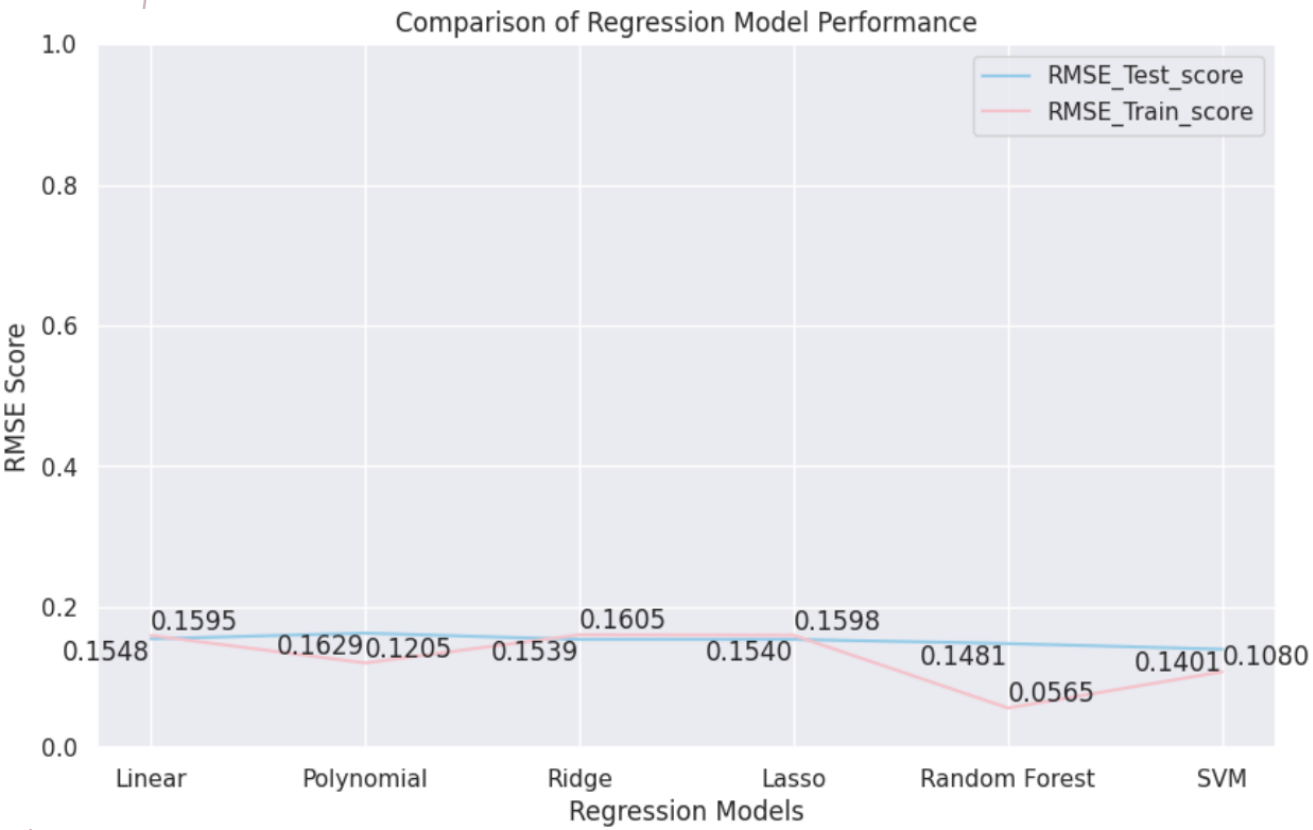
METRICS EVALUATION

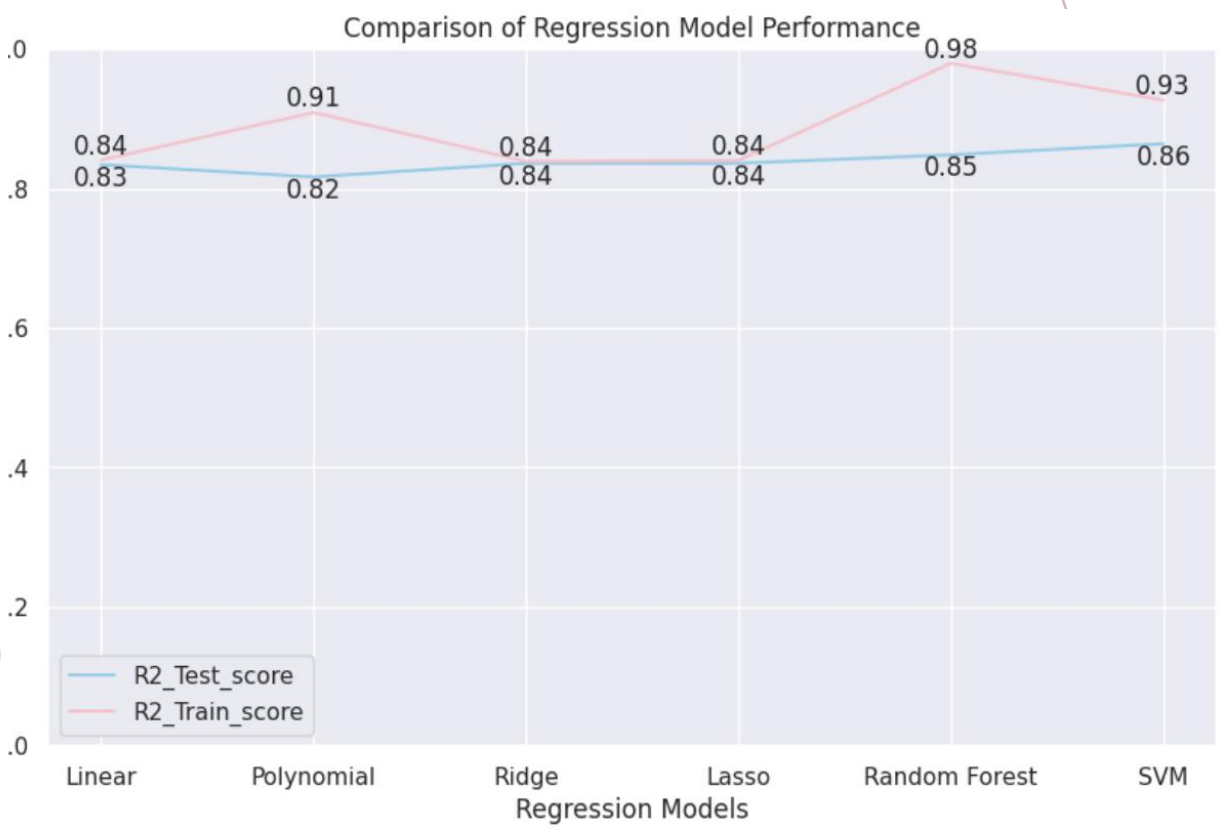
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{X}_i)^2}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum_{i=1}^{n}(X_i - \hat{X}_i)^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

	Linear	Polynomial	Ridge	Lasso	Random Forrest	SVM
Train RMSE	0.1595	0.1205	0.1604	0.1597	0.0565	0.1080
Test RMSE	0.1548	0.1628	0.1539	0.1539	0.1493	0.1400
Train R2	0.8346	0.9094	0.8393	0.8407	0.9800	0.9272
Test R2	0.8412	0.8169	0.8365	0.8365	0.8460	0.8616

MODELS PERFORMANCE





SUMMARY

In conclusion, the project was performed by following a checklist of steps for implementing an ML task. We chose several regression models and evaluated their metrics. As a result, the close R2 scores between training and testing sets indicate strong predictive performance across models, particularly the SVM, which demonstrates notable generalization




THANK YOU

Pavlo Poliuha

Chen Junjie

Pablo Mollá