House Price Prediction — Regularised Regression

Author : K. Karthik Kumar Reddy
Program : SAP labs COHORT3
Date : 17 January 2026

Subjective Answers

Question 1

What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

ANS :

- Optimal alpha definition : The optimal alpha (regularisation strength) is the value selected by cross-validation that minimises validation error (typically MSE). For Ridge/Lasso, alpha controls penalty strength : larger alpha → stronger shrinkage.
- Effect of doubling alpha : Doubling alpha increases shrinkage. For Ridge, coefficients are uniformly pulled closer to zero but typically remain non-zero; model variance decreases and bias increases, often increasing training error but improving generalisation in overfit cases. For Lasso, doubling alpha can zero out more coefficients (feature selection), reducing model complexity and possibly removing weaker predictors entirely. The net effect is fewer features and sparser models for Lasso, and smaller coefficients for Ridge.
- Which predictors become most important after doubling : Predictors with the largest absolute coefficients before doubling are likeliest to remain important after doubling. Under Lasso, only a small subset with strong signal-to-noise will survive; under Ridge, most original predictors stay but with reduced magnitudes. Exact variable names depend on the dataset and must be read from the model outputs (see `model_summary.json`).

Question 2
You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

ANS :

- Choice principle : Prefer the model that best balances predictive accuracy and interpretability on held-out data. If Lasso achieves similar or better CV error and yields a sparse model, choose Lasso for easier interpretation and deployment. If Lasso underperforms or removes predictors known to be important, choose Ridge for stability and better coefficient shrinkage without aggressive variable exclusion.
- Practical considerations : Use Lasso when variable selection is desirable and multicollinearity exists; use Ridge when many predictors contribute small effects or when you want coefficient stability. The final decision must be supported by cross-validated RMSE and business needs.

Question 3
After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model

excluding the five most important predictor variables. Which are the five most important predictor variables now?

ANS :

- Retrain Lasso excluding the original top-5 predictors (identified from the first Lasso). Then rank coefficients by absolute value to get the new top-5. This yields a robust fallback model that relies on the next-most-informative features.

- Interpretation : The new top-5 will typically include predictors that were moderately important previously or predictors correlated with the removed ones. Exact names and coefficients are produced by the retraining step in `run_house_model.py` and saved into `model_summary.json` under `lasso_retrained_top`.

Question 4

How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?-

ANS :

Methods to ensure robustness / generalisability :

- Use cross-validation (k-fold or repeated) to estimate out-of-sample performance and select hyperparameters.

- Keep a hold-out test set for final evaluation that is never used during training or hyperparameter tuning.

- Apply regularisation (Ridge/Lasso/ElasticNet) to reduce variance and prevent overfitting.

- Use domain-aware feature engineering and avoid leakage (do not include future or derived variables that won't be available at scoring time).

- Detect and handle data imbalance, missingness, and distribution shifts (retrain or adapt when distributions change).

- Prefer simpler models when they reach similar performance (Occam's razor).

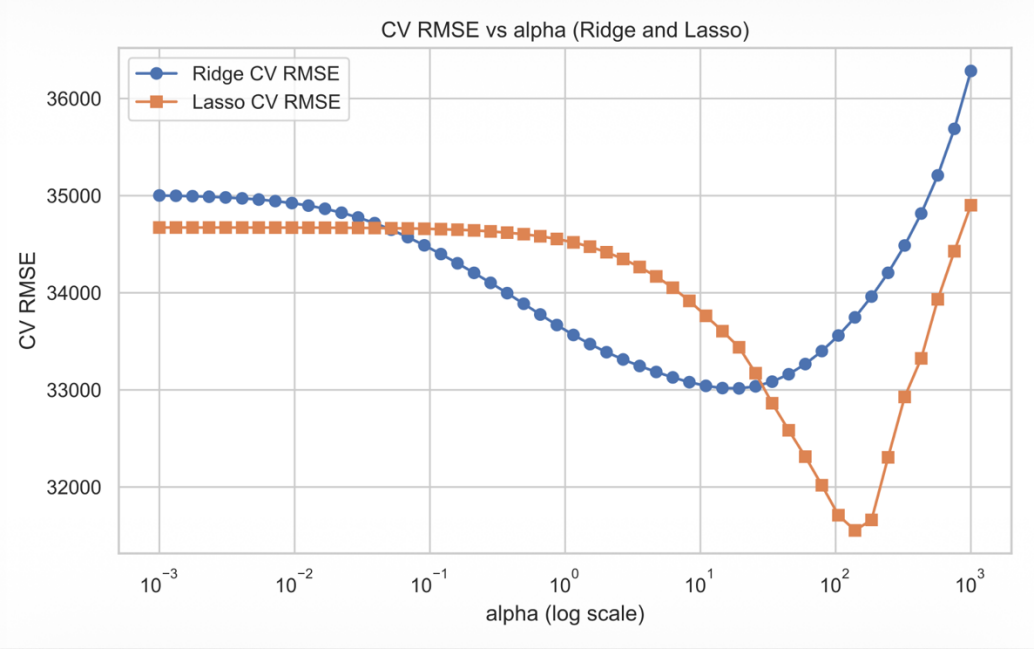
- Validate model stability across subgroups and time slices to confirm consistent behaviour.

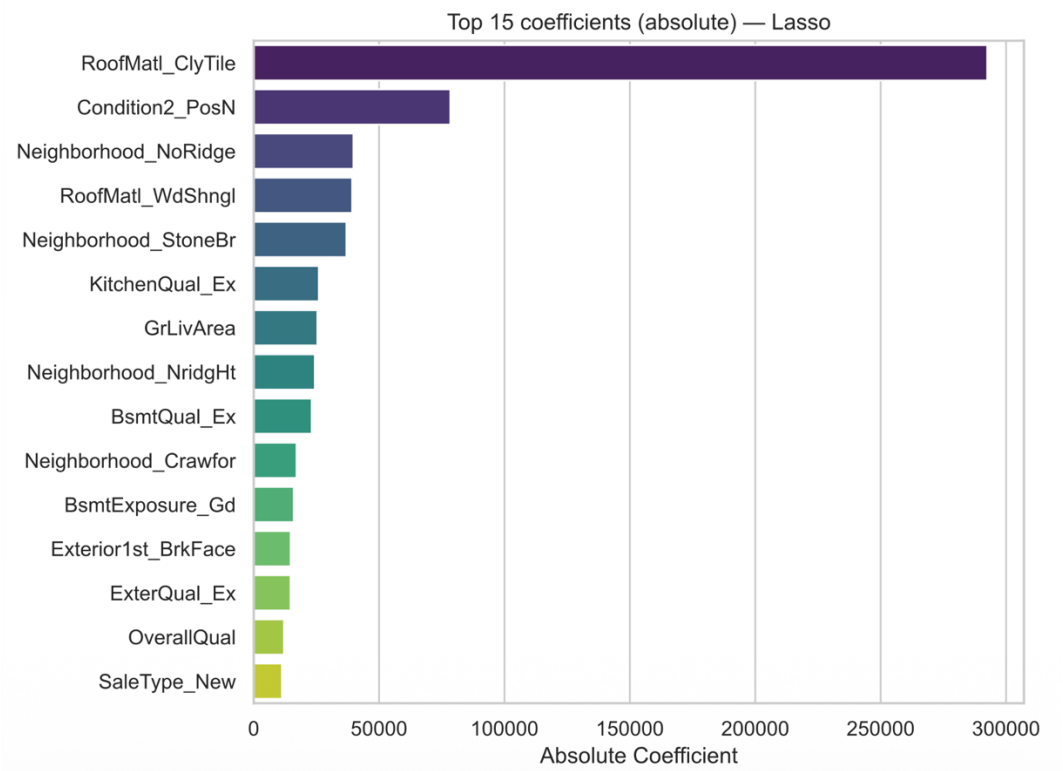
- Implications for accuracy : A robust, generalisable model often sacrifices some training accuracy (higher bias) in exchange for lower test error (lower variance). The overall goal is better out-of-sample accuracy rather than perfect in-sample fit. Regularisation and simpler models reduce variance and improve reliability on new data, which is more valuable for deployment.

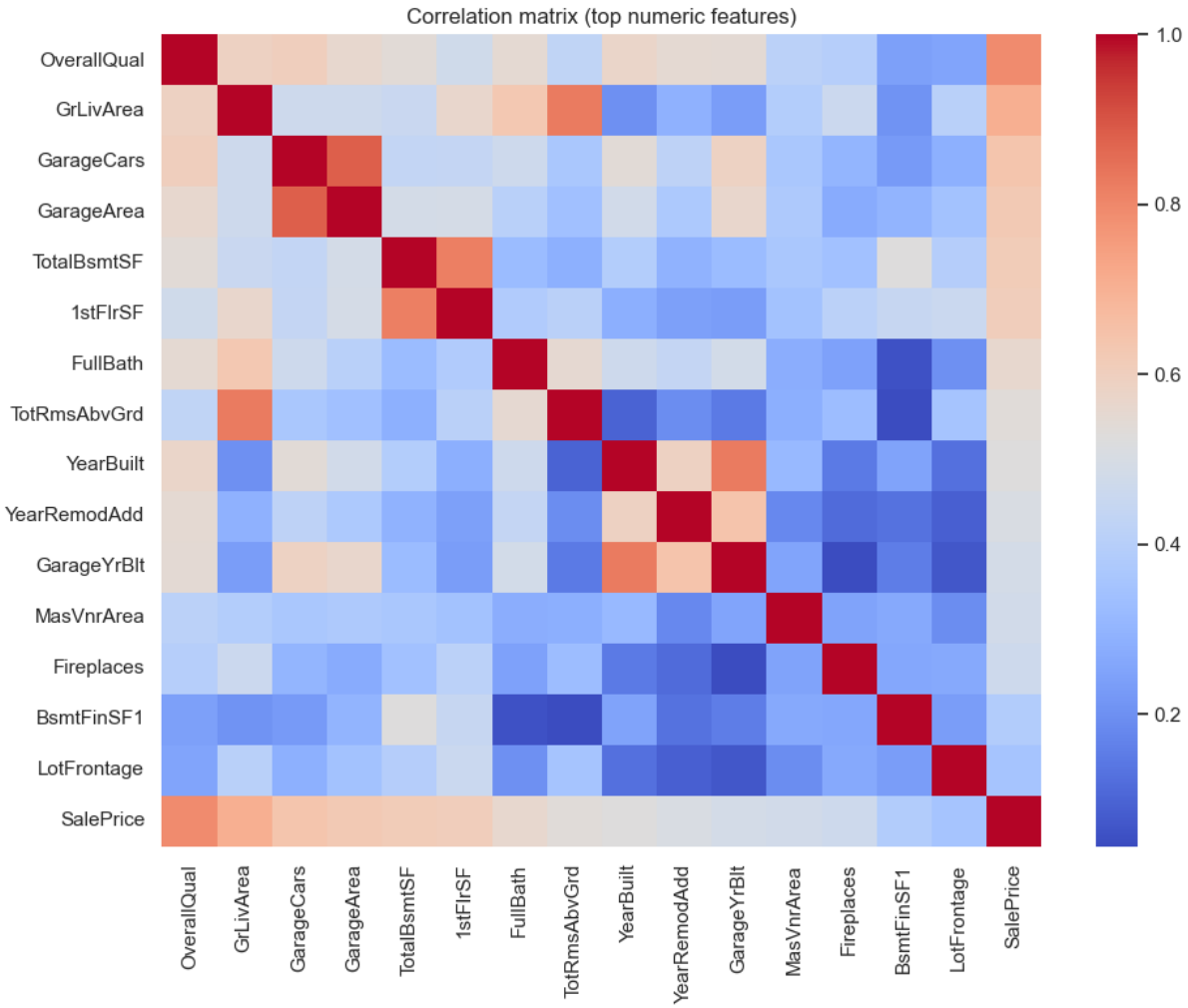
DATA ANALYSIS AND ILLUSTRATION FIGURES

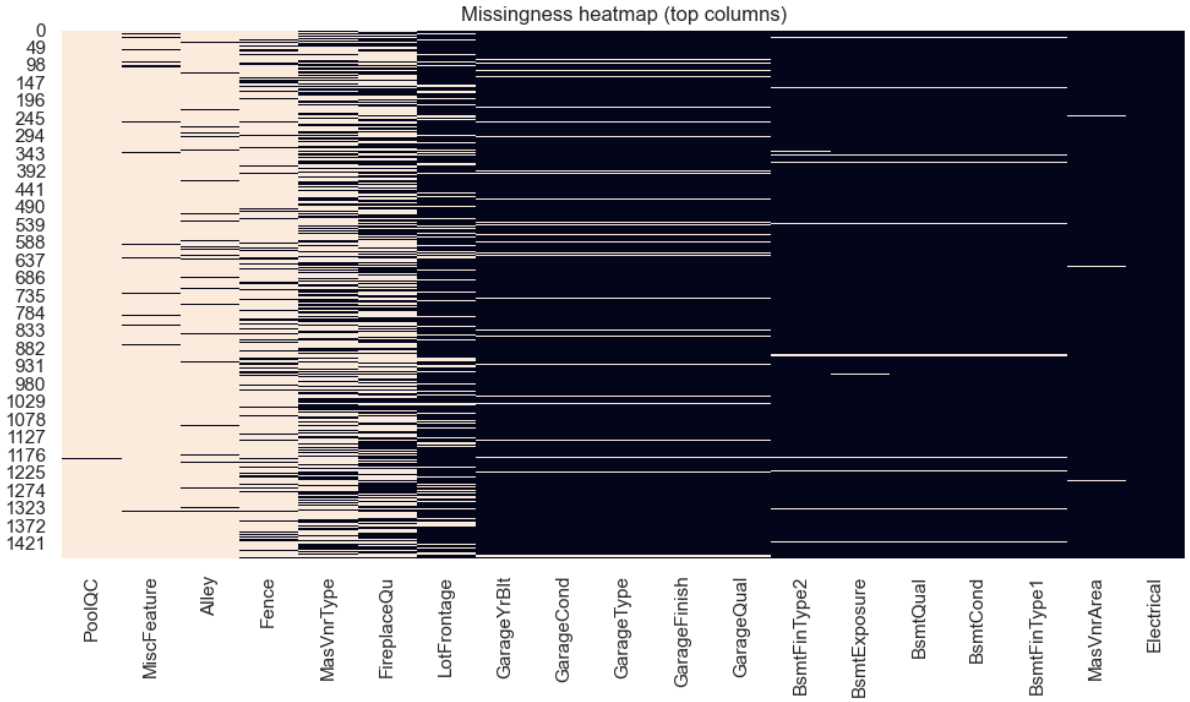
EDA Figure Example :


SalePrice distribution
Count
SalePrice
0
25
50
75
100
125
150
175
0
100000
200000
300000
400000
500000
600000
700000


CV RMSE vs alpha (Ridge and Lasso)
Ridge CV RMSE
Lasso CV RMSE
CV RMSE
alpha (log scale)
32000
33000
34000
35000
36000
10^{-3}
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}
10^{3}


Top 15 coefficients (absolute) — Lasso
RoofMatl_ClyTile
Condition2_PosN
Neighborhood_NoRidge
RoofMatl_WdShngl
Neighborhood_StoneBr
KitchenQual_Ex
GrLivArea
Neighborhood_NridgHt
BsmtQual_Ex
Neighborhood_Crawfor
BsmtExposure_Gd
Exterior1st_BrkFace
ExterQual_Ex
OverallQual
SaleType_New
0
50000
100000
150000
200000
250000
300000
Absolute Coefficient


Correlation matrix (top numeric features)
OverallQual
GrLivArea
GarageCars
GarageArea
TotalBsmtSF
1stFlrSF
FullBath
TotRmsAbvGrd
YearBuilt
YearRemodAdd
GarageYrBlt
MasVnrArea
Fireplaces
BsmtFinSF1
LotFrontage
SalePrice
1.0
0.8
0.6
0.4
0.2


Missingness heatmap (top columns)
0
49
98
147
196
245
294
343
392
441
490
539
588
637
686
735
784
833
882
931
980
1029
1078
1127
1176
1225
1274
1323
1372
1421
PoolQC
MiscFeature
Alley
Fence
MasVnrType
FireplaceQu
LotFrontage
GarageYrBlt
GarageCond
GarageType
GarageFinish
GarageQual
BsmtFinType2
BsmtExposure
BsmtQual
BsmtCond
BsmtFinType1
MasVnrArea
Electrical

One-page Summary

Ridge CV RMSE: 33791.44670925614
Lasso CV RMSE: 31904.08949706496
Ridge alpha: 14.563484775012444
Lasso alpha: 138.9495494373136