ZHANG ZHE A1775543

BIG DATA ANALYSIS & PROJECT

HOUSE PRICES: ADVANCED REGRESSION TECHNIQUES

GOAL

- Predict the sales price for each house.
- Required to use the FIRST (not random) 70% of the training samples to form a "small training set" and the remaining 30% of the training samples to form a "small validation set". Then you will develop your algorithm by training your model on the small training set and evaluating on the small validation set.
- Run the model on the test set to obtain the estimation

METRIC

- Results are evaluated on Root-Mean-Squared-Error (RMSE) between the logarithm of the predicted value and the logarithm of the observed sales price. (Taking logs means that errors in predicting expensive houses and cheap houses will affect the result equally.)

DATA COLLECTION

- The quantity & quality of your data dictate how accurate our model is
- The outcome of this step is generally a representation of data which we will use for training
- Using pre-collected data, by way of datasets from Kaggle, UCI, etc., still fits into this step

DATA PREPARATION

- Clean that which may require it (remove duplicates, correct errors, deal with missing values, normalization, data type conversions, etc.)
- Visualize data to help detect relevant relationships between variables or class imbalances (bias alert!), or perform other exploratory analysis
- Split into training and evaluation sets

CHOOSE A MODEL

- Different algorithms are for different tasks; choose the right one

TRAIN THE MODEL

- The goal of training is to answer a question or make a prediction correctly as often as possible

EVALUATE THE MODEL

- Uses some metric or combination of metrics to "measure" objective performance of model
- Test the model against previously unseen data
- This unseen data is meant to be somewhat representative of model performance in the real world, but still helps tune the model (as opposed to test data, which does not)

PARAMETER TUNING

- This step refers to hyperparameter tuning, which is an "artform" as opposed to a science
- Tune model parameters for improved performance
- Simple model hyperparameters may include: number of training steps, learning rate, initialization values and distribution, etc.

MAKE PREDICTIONS

- Using further (test set) data which have, until this point, been withheld from the model (and for which class labels are known), are used to test the model; a better approximation of how the model will perform in the real world

DATA COLLECTION

- Given by the Kaggle

DATA PREPARATION

- Data Split: using iloc
- Review data: sns.distplot(data_train['SalePrice'])
- Check the heatmap: sns.heatmap(data)
- Handling abnormal data: data.drop
- Handling missing data
- Convert Series to dummy codes: get_dummies

BASIC MODELING & EVALUATION

- Linear Regression
- Ridge
- Lasso
- Random Forrest
- Gradient Boosting Tree
- Support Vector Regression
- Linear Support Vector Regression
- ElasticNet
- Stochastic Gradient Descent
- BayesianRidge
- KernelRidge
- ExtraTreesRegressor
- XgBoost

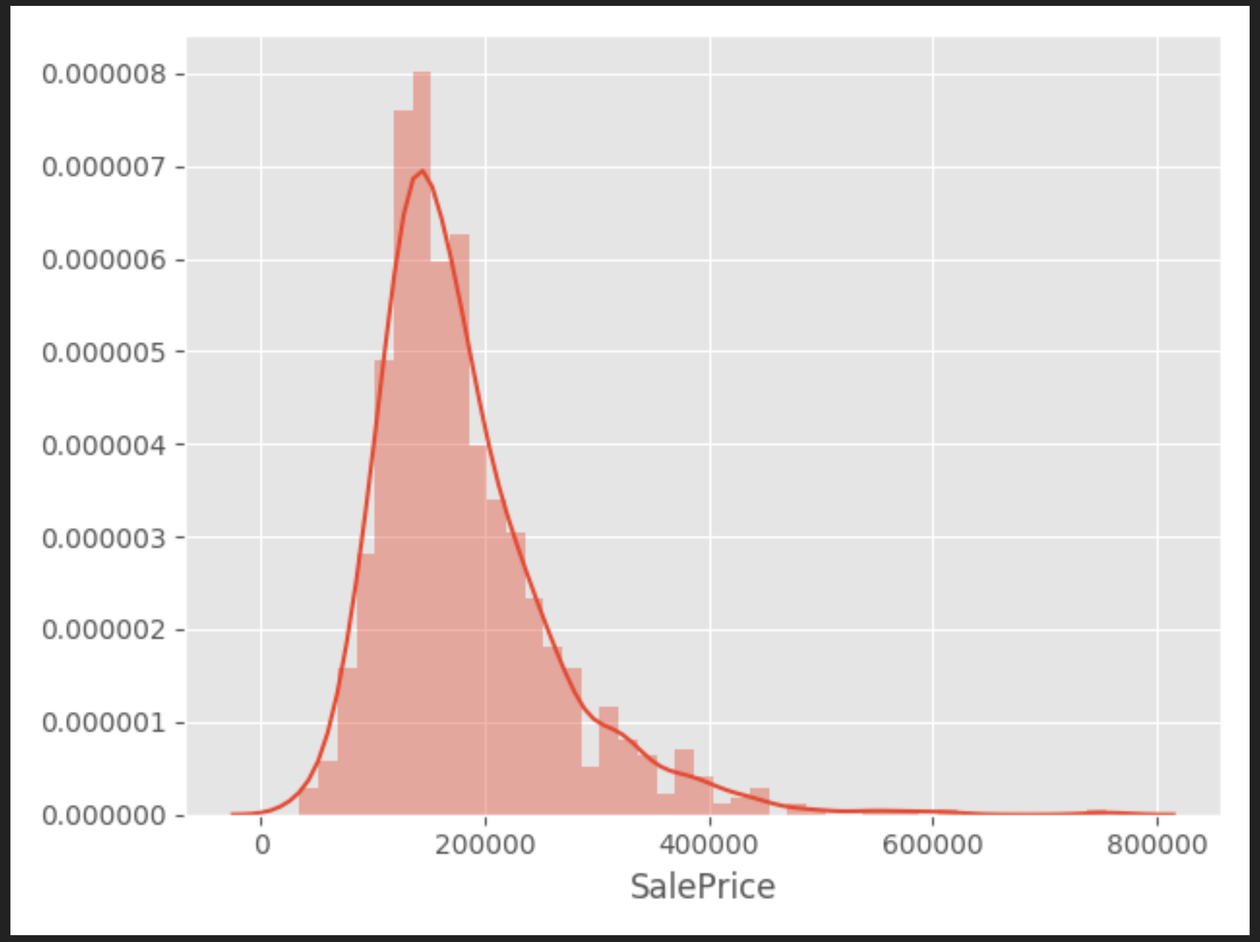
PARAMETER TUNING

- lasso = Lasso(alpha=0.0005,max_iter=10000)
- ridge = Ridge(alpha=60)
- svr = SVR(gamma= 0.0004,kernel='rbf',C=13,epsilon=0.009)
- ker = KernelRidge(alpha=0.2 ,kernel='polynomial',degree=3 , coef0=0.8)
- ela = ElasticNet(alpha=0.005,l1_ratio=0.08,max_iter=10000)
- bay = BayesianRidge()

CURRENT SCORE

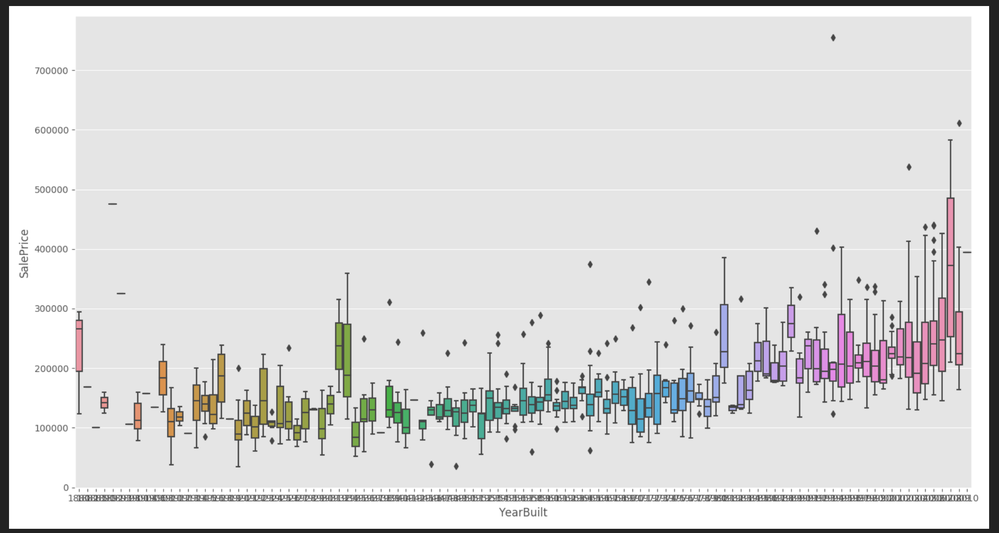
- 0.13727146176659805

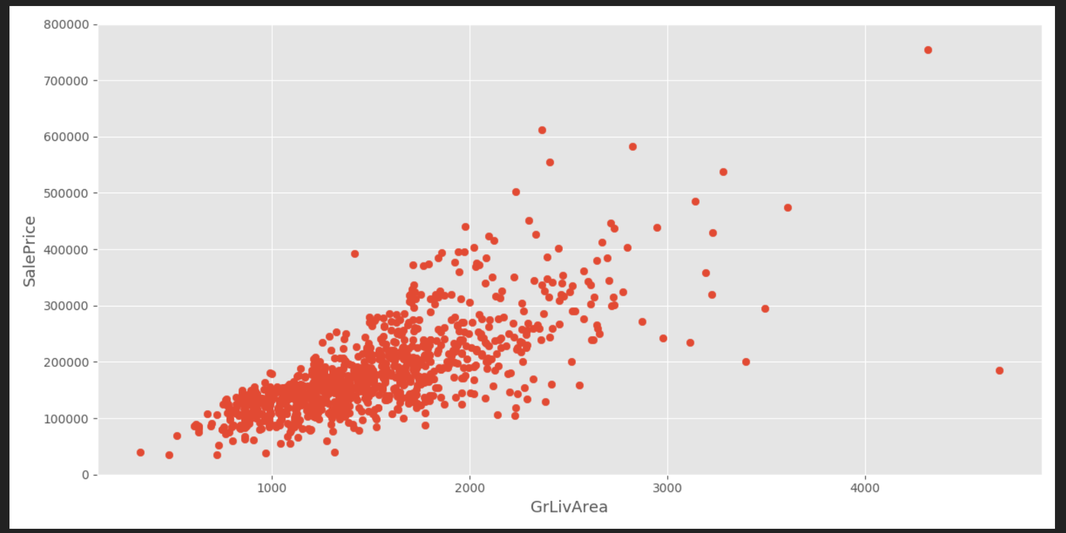
DATA DISTRIBUTION



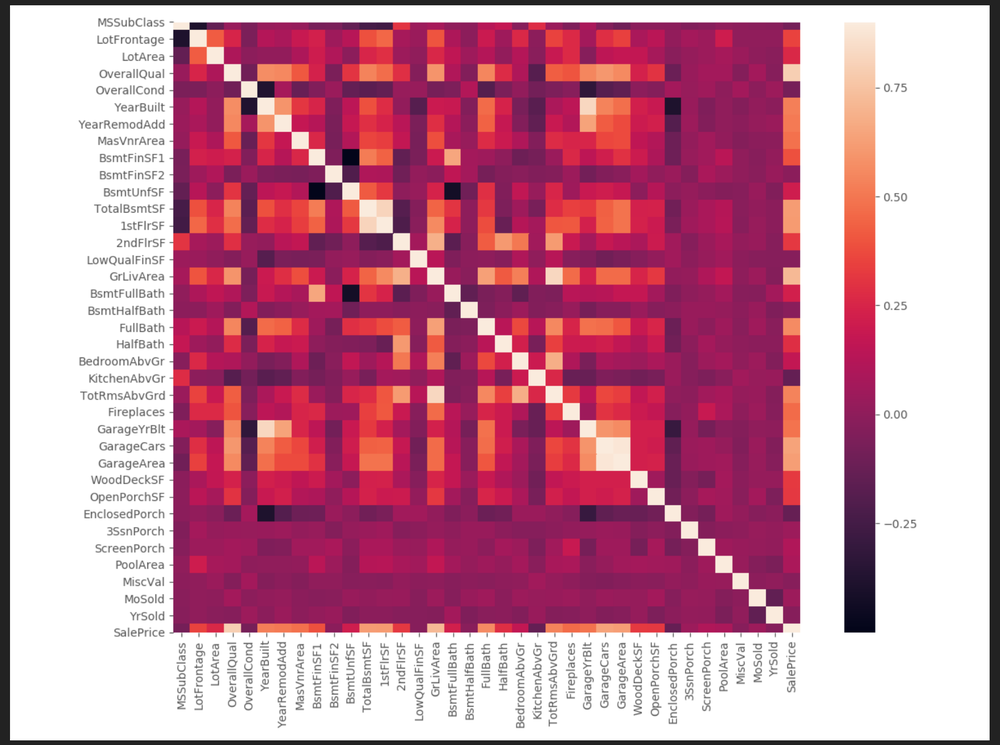
count	1021.00
mean	181701.22
std	79892.87
min	34900.00
25%	130000.00
50%	163500.00
75%	215000.00
max	755000.00

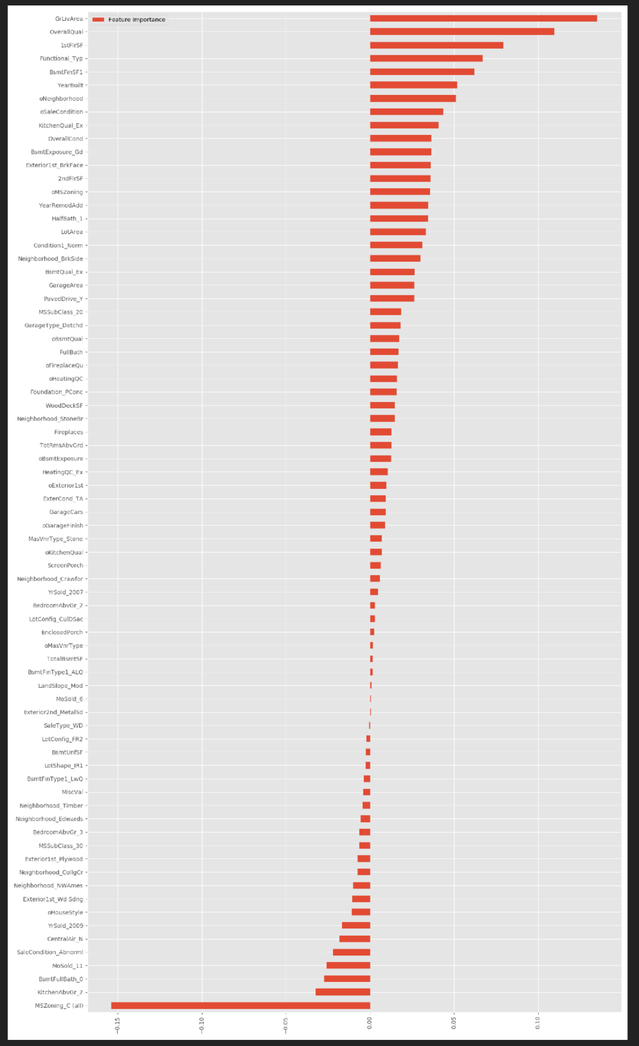
EXPLORATORY VISUALIZATION





FEATURE IMPORTANCE





CONCLUSION

- Currently using Lasso，Ridge，SVR，Kernel Ridge，ElasticNet，Bayesian Ridge as first layer, and Kernel Ridge for the second layer. Still working on finding better result.
- It’s worth to spend time on featuer engineering, maybe I will improve this part if time allowed.
- it’s better to spend more time on data visualization.

REFERENCE

- https://www.kaggle.com/massquantity/all-you-need-is-pca-lb-0-11421-top-4/data
- https://www.kaggle.com/serigne/stacked-regressions-top-4-on-leaderboard
- https://www.kaggle.com/neviadomski/how-to-get-to-top-25-with-simple-model-sklearn