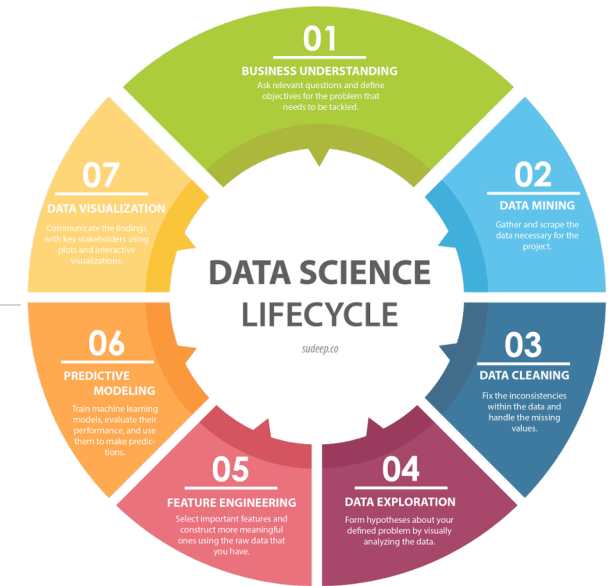


# Ames Housing Project: Advanced regression techniques for predictive model

Mori Esam – Jan 2020



# Build a model to :

Predict the price of a house at sale

Target : Real estate agencies, mortgage brokers, ...

Test key amenities/features that influence house's value

Target : Group 1 + Architects, interior/exterior designers, contractors

# Overview

Identify key metrics  
that actually matter



Neighborhood   Location		Home Size   Usable Space		Features		Age   Condition	
Nominal	neighborhood ms_subclass ms_zonning pid	Discrete	full_bath bedroom_abvgr kitchen_abvgr totrms_abvgrd garage_cars	Nominal	bldg_type house_style foundation functional land_countour lot_config roof_style roof_matl mas_vnr_type exterior_1st exterior_2nd Fence heating utilities central_air garage_type	Discrete	year_built year_remod/add mo_sold yr_sold garage_yr_blt
		Continuous	bsmtfin_sf_1 bsmtfin_sf_2 bsmt_unf_sf total_bsmt_sf 1st_flr_sf 2nd_flr_sf gr_liv_area garage_area wood_deck_sf open_porch_sf lot_area lot_frontage mas_vnr_area			Ordinal	overall_qual exter_qual exter_cond bsmt_qual bsmt_cond fireplace_qu garage_qual kitchen_qual heating_qc
						Nominal	condition_1 sale_type

# Implementation Process

## Data Cleaning

- Consistency
- Missing Data
- Categorical Data

Renaming

Null → No Features

Null → Mode or Median

Conversion to numeric dummies  
**NO** - use *directional attributes*

## EDA

- Identify relevant var(s)
- Linear Relationship
- Normalize Distribution

Outline Strategy : meaningful data

Attributes with  $> 0.5$  linear corr.

Log Transform SalePrice

Outliers

**Do not drop** - avoid forcing  
model to appear less variable  
than it is in reality.

## Preprocessing Modeling

- Set up matrix & target
- Choose the best model
- Instantiate | Fit | Cross Validation  
    > *Feature engineering*
- Generate Prediction
- Evaluate

MLR > LASSO > Ridge

Interaction term improved  $R^2$

Goal

*test  $R^2 \sim 0.9$*

*MAE  $\sim 0$*

*MSE  $\sim 0$*

*RMSE  $\sim \$10K$*

# Result : MLR Model

## Regression Metrics

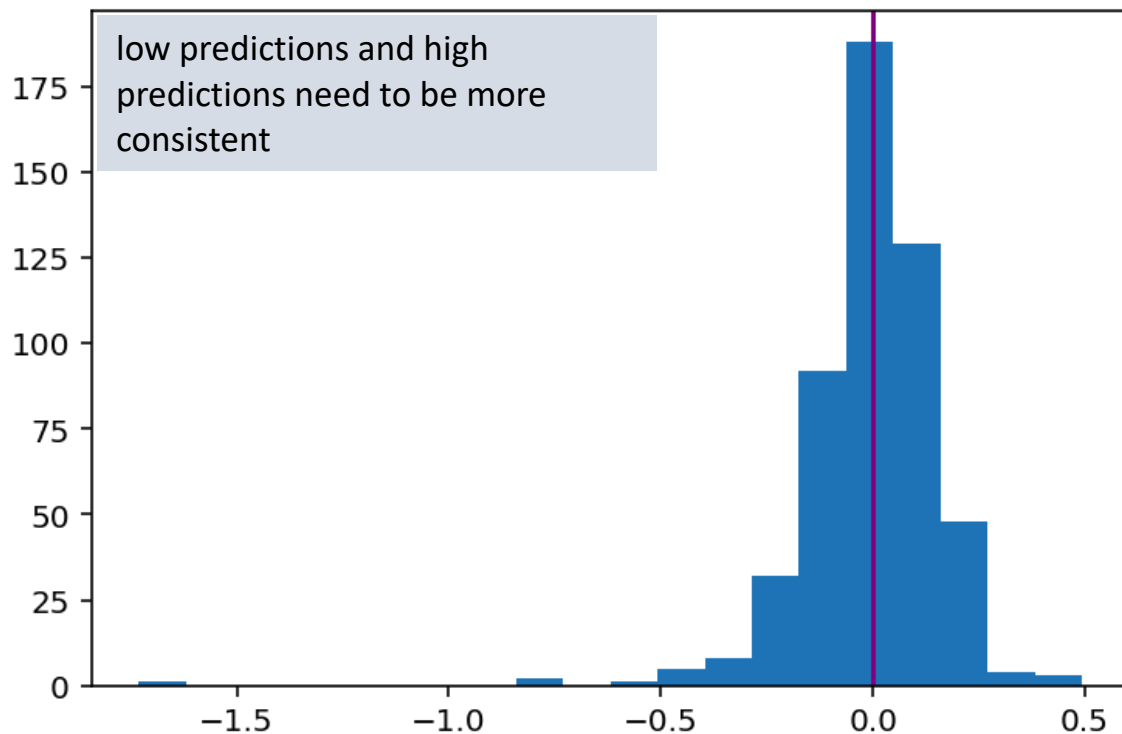
***R*<sup>2</sup> Score : 0.853**

**MAE : 0.107**

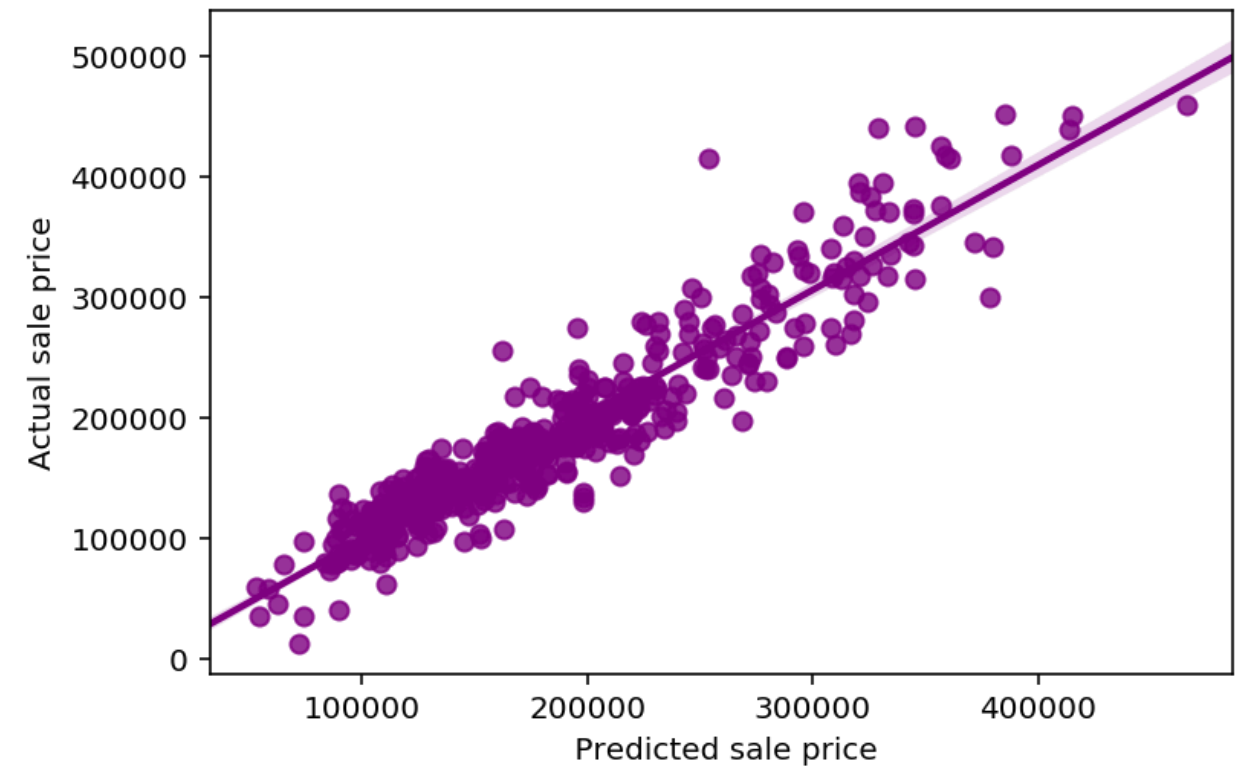
**MSE : 0.008**

**RMSE : \$27K**

**Residual (sale price – actual vs predicted)**



**Predicted vs actual price**



# Conclusion and Recommendations:

As of now **85.3%** of the variability in sale price can be explained by this model, indicating that the model is just right. In other words, model can generalize from train/test data to predict the house value, with **+/- \$27K** price error. Nevertheless, this model can be used for the following purposes:

## **Prediction -**

This model can be used for any dataset that includes similar attributes on house features to predict the property's selling price .

## **Inference -**

This model can be used to outline and test some of the most important factors/features that influence house's value.

## **Optimize for accuracy -**

Improve model by leveraging feature engineering to include interesting features like Neighborhood, house style and materials.