**Contents** :

1. [**Introduction**](#_Introduction:)
2. [**Dataset**](#_Dataset:)
3. [**Data Wrangling**](#_Data_Wrangling:)
4. [**Exploratory Data Analysis**](#_Exploratory_Data_Analysis)
5. [**Modeling**](#_Modeling:)

# [Introduction](#First_heading)

Ames Housing dataset has 79 explanatory variables. This is from Kaggle. This dataset is a clean version. Each row in the dataset describes the characteristics of a house.

The goal of this project is to:

* Predict the SalePrice, given these features.
* Our models are evaluated on the Root-Mean-Squared-Error (RMSE)

# Dataset

Data fields are SalePrice, Zoning, Previous Pricing information, LotArea, YearBuilt, Garage, Bedroom, Bathroom, Kitchen, Basement and other details.



MSSubClass: Identifies the type of dwelling involved in the sale.

MSZoning: Identifies the general zoning classification of the sale.

LotFrontage: Linear feet of street connected to property

LotArea: Lot size in square feet

Street: Type of road access to property

Alley: Type of alley access to property

LotShape: General shape of property

LandContour: Flatness of the property

Utilities: Type of utilities available

LotConfig: Lot configuration

LandSlope: Slope of property

Neighborhood: Physical locations within Ames city limits

BldgType: Type of dwelling

HouseStyle: Style of dwelling

OverallQual: Rates the overall material and finish of the house

OverallCond: Rates the overall condition of the house

YearBuilt: Original construction date

YearRemodAdd: Remodel date (same as construction date if no remodeling or additions)

RoofStyle: Type of roof

RoofMatl: Roof material

Exterior1st: Exterior covering on house

Exterior2nd: Exterior covering on house (if more than one material)

MasVnrType: Masonry veneer type

MasVnrArea: Masonry veneer area in square feet

ExterQual: Evaluates the quality of the material on the exterior

ExterCond: Evaluates the present condition of the material on the exterior

Foundation: Type of foundation

BsmtQual: Evaluates the height of the basement

BsmtCond: Evaluates the general condition of the basement

BsmtExposure: Refers to walkout or garden level walls

BsmtFinType1: Rating of basement finished area

BsmtFinSF1: Type 1 finished square feet

BsmtFinType2: Rating of basement finished area (if multiple types)

BsmtFinSF2: Type 2 finished square feet

BsmtUnfSF: Unfinished square feet of basement area

TotalBsmtSF: Total square feet of basement area

HeatingQC: Heating quality and condition

CentralAir: Central air conditioning

Electrical: Electrical system

1stFlrSF: First Floor square feet

2ndFlrSF: Second floor square feet

LowQualFinSF: Low quality finished square feet (all floors)

GrLivArea: Above grade (ground) living area square feet

BsmtFullBath: Basement full bathrooms

BsmtHalfBath: Basement half bathrooms

FullBath: Full bathrooms above grade

HalfBath: Half baths above grade

Bedroom: Bedrooms above grade (does NOT include basement bedrooms)

Kitchen: Kitchens above grade

KitchenQual: Kitchen quality

Functional: Home functionality (Assume typical unless deductions are warranted)

Fireplaces: Number of fireplaces

FireplaceQu: Fireplace quality

GarageType: Garage location

GarageYrBlt: Year garage was built

GarageFinish: Interior finish of the garage

GarageCars: Size of garage in car capacity

GarageArea: Size of garage in square feet

GarageQual: Garage quality

GarageCond: Garage condition

PavedDrive: Paved driveway

WoodDeckSF: Wood deck area in square feet

OpenPorchSF: Open porch area in square feet

EnclosedPorch: Enclosed porch area in square feet

3SsnPorch: Three season porch area in square feet

ScreenPorch: Screen porch area in square feet

PoolArea: Pool area in square feet

PoolQC: Pool quality

Fence: Fence quality

MiscFeature: Miscellaneous feature not covered in other categories

MiscVal: $Value of miscellaneous feature

MoSold: Month Sold (MM)

YrSold: Year Sold (YYYY)

SaleType: Type of sale

SaleCondition: Condition of sale

# Data Wrangling

Of the 79 explanatory variables, various steps were taken for data wrangling.

Dropping features & Log transformation

* SalePrice & ID is dropped from the feature set as SalePrice is the target variable that needs to be predicted.
* Since SalePrice is highly skewed and not normally distributed, log1p transformation is applied to improve prediction accuracy

## Converting few numerical features to string

MSSubClass, & YrSold and MoSold are converted to string

## Imputing missing values

## Creating a separate category for missing values for features that makes sense

Some of the features where the values are missing makes sense. For example, houses that don’t have a pool, basement, garage is replaced with “None”.

Other numerical features like

## Categorical

Imputing missing values for features like Functional, Electrical, KitchenQual, MSSubClass, MSZoning with their mode

LotFrontage is imputed with lot’s median by Neighborhood

## Numerical

## Fix skewed features by normalizing it & using Box Cox Transformation

Features that are skewed, whose skew value is greater than 0.5 are transformed using Box Cox transformation.

## Creating other interesting derived features

## 

* YrBltAndRemod : Year built + Year remodeled
* TotalSF: TotalBsmtSF + 1stFlrSF+ 2ndFlrSF
* Total\_Bathrooms : FullBath + 0.5 \* HalfBath + BsmtFullBath+ 0.5 \* BsmtHalfBath
* Total\_porch\_sf : OpenPorchSF + 3SsnPorch + EnclosedPorch + ScreenPorch+ WoodDeckSF
* Creating binary features to identify whether a house has certain features: Haspool, has2ndfloor, hasgarage, hasbsmt, hasfireplace

## Creating dummy variables for all the features

This was done for all the features that needed it

## Outliers

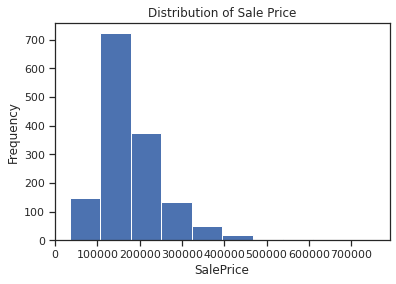
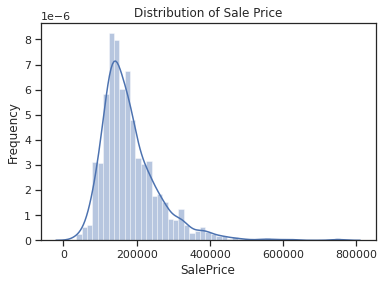
Based on eda learnings, manually removed some of the rows

Cross Validation

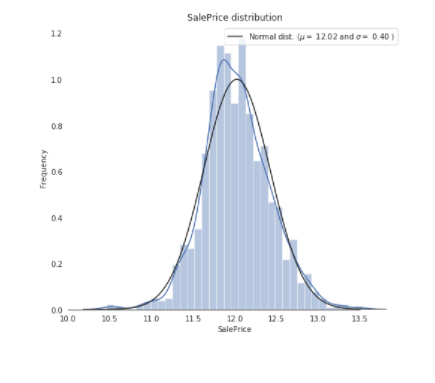
Using 12 fold cross validation for testing

# Exploratory Data Analysis

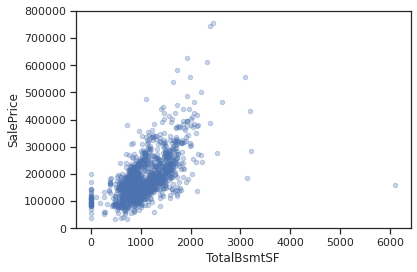
SalePrice is skewed and not normally distributed. Below is the distribution :

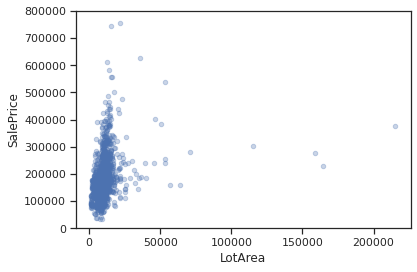
 

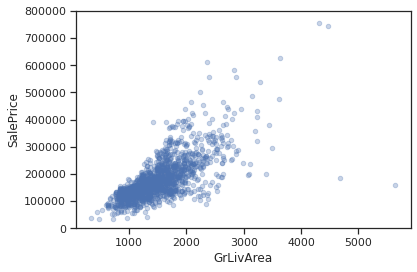
To fix its skewness, log1P transformation was applied and post transformation, the plot looks like below:



**Features that are highly correlated with SalePrice :**

****

****

****

# Modeling

Tried a few models and selected stacked models as that has the best rmse value compared to others.

|  |  |
| --- | --- |
| **Model** | **Root Mean Square Error** |
| Lasso | 0.1201 |
| LightGBM | 0.1219 |
| Elastic Net Regression | 0.1196 |
| SVR | 0.1184 |
| LightGBM | 0.1219 |
| Gradient Boosting | 0.1232 |
| XGBoost | 0.1198 |

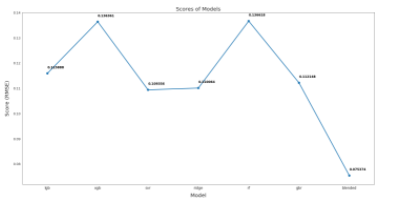
Stacking : Trained StackingCV Regressor to use a blended model approach and was optimized using XGboost.

For predictions, below blended approach was used to make the predictions :

0.1 \* elastic\_model+ 0.05 \* lasso\_model+ 0.1 \* ridge\_model+ 0.1 \* svr\_model + 0.1 \* gbr\_model+ 0.15 \* xgb\_model\_full\_data+ 0.1 \* lgb\_model+ 0.3 \* stack\_gen\_model

RMSE : 0.0616

We can observe from the graph above that the blended model outperforms the other models, with an RMSLE of 0.075. This is the model that was used to make predictions.



*Related sources : https://www.kaggle.com/lavanyashukla01/how-i-made-top-0-3-on-a-kaggle-competition*