IBM Supervised Machine Learning Regression Assignment

By

Ahmed Shahriar Sakib

Introduction:

The objective of this project is to perform exploratory data analysis on **Ames Housing Price Dataset** and build regression machine learning models to predict housing prices.

Main Objective

- Check for null values and perform imputation
- Check for duplicated rows or features
- Perform scaling and normalization on features
- Build machine learning models and optimize them for regression task
- Perform model evaluation

Dataset:

Source

- Kaggle Dataset URL: https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Exploratory Data Analysis

Dataset dimension:

- **Train**: 1460 rows, and 81 columns; **Test**: 1459 rows, and 80 columns

Data Types:

- Train Dataset :

Data Type	Count
object	43
int64	35
float64	3

- Test Dataset :

Data Type	Count
object	43
int64	26
float64	11

Brief Description of Some Features

Here's a brief version of what you'll find in the data description file -

- SalePrice - the property's sale price in dollars. This is the target variable that you're trying to predict.
- MSSubClass: The building class
- MSZoning: The general zoning classification
- LotFrontage: Linear feet of street connected to property
- LotArea: Lot size in square feet
- Street: Type of road access
- Alley: Type of alley access
- 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
- Condition1: Proximity to main road or railroad
- Condition2: Proximity to main road or railroad (if a second is present)
- BldgType: Type of dwelling
- HouseStyle: Style of dwelling
- OverallQual: Overall material and finish quality
- OverallCond: Overall condition rating
- YearBuilt: Original construction date
- YearRemodAdd: Remodel date
- 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: Exterior material quality
- ExterCond: Present condition of the material on the exterior
- Foundation: Type of foundation
- BsmtQual: Height of the basement
- BsmtCond: General condition of the basement
- BsmtExposure: Walkout or garden level basement walls
- BsmtFinType1: Quality of basement finished area
- BsmtFinSF1: Type 1 finished square feet
- BsmtFinType2: Quality of second finished area (if present)

- BsmtFinSF2: Type 2 finished square feet
- BsmtUnfSF: Unfinished square feet of basement area
- TotalBsmtSF: Total square feet of basement area
- Heating: Type of heating
- 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: Number of bedrooms above basement level
- Kitchen: Number of kitchens
- KitchenQual: Kitchen quality
- TotRmsAbvGrd: Total rooms above grade (does not include bathrooms)
- Functional: Home functionality rating
- 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

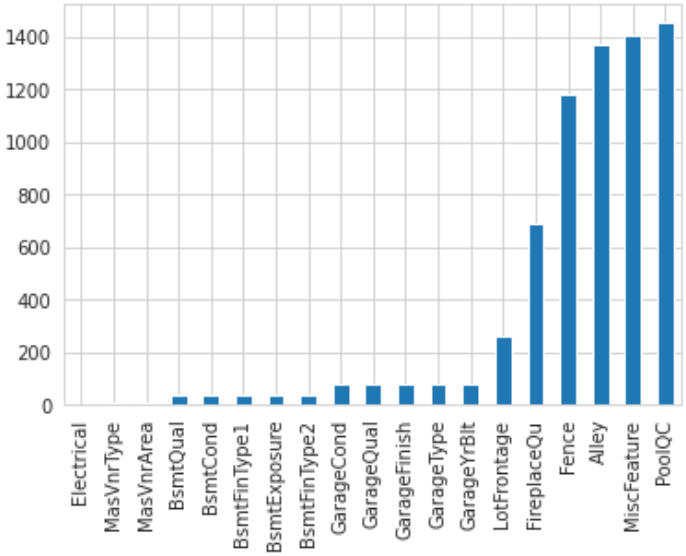
Features Summary (Train Dataset)

Qualitative features: 43

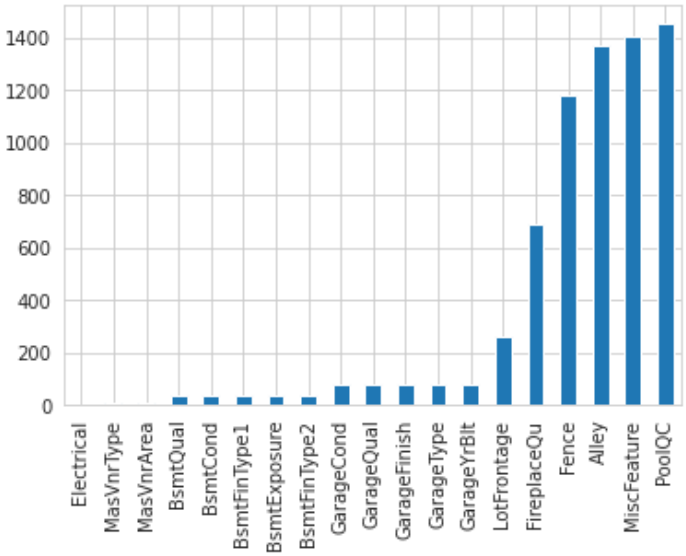
Quantitative features: 36

Missing Value

Train Dataset



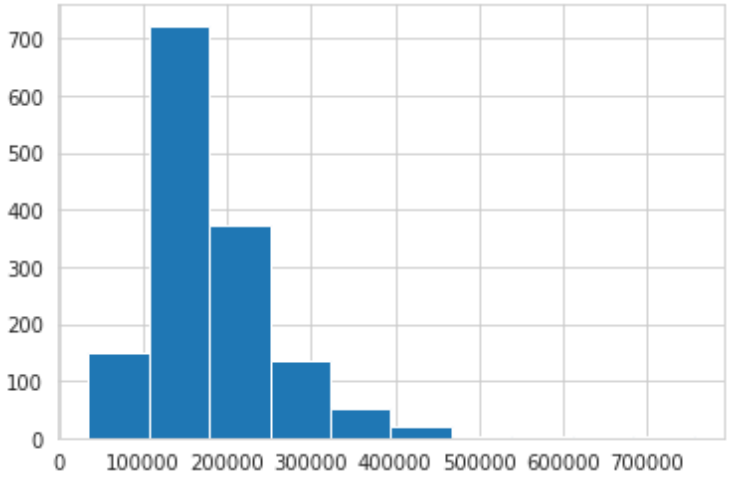
Test Dataset



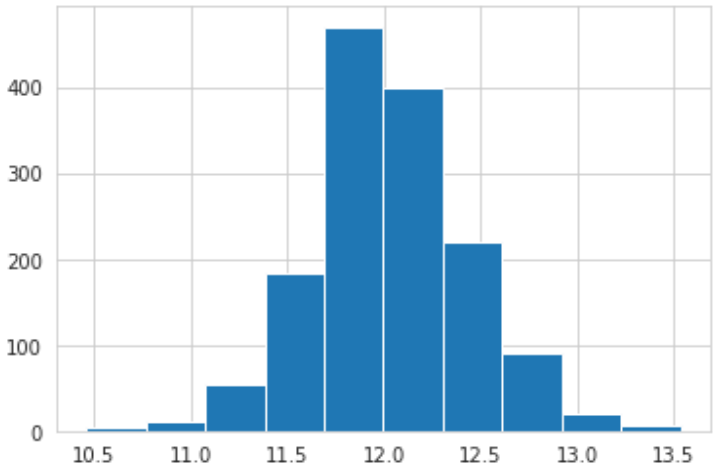
Normality Check

Skewed Target (‘SalePrice’”)

Before Normalization (skewness : 1.883) -



After normalization –

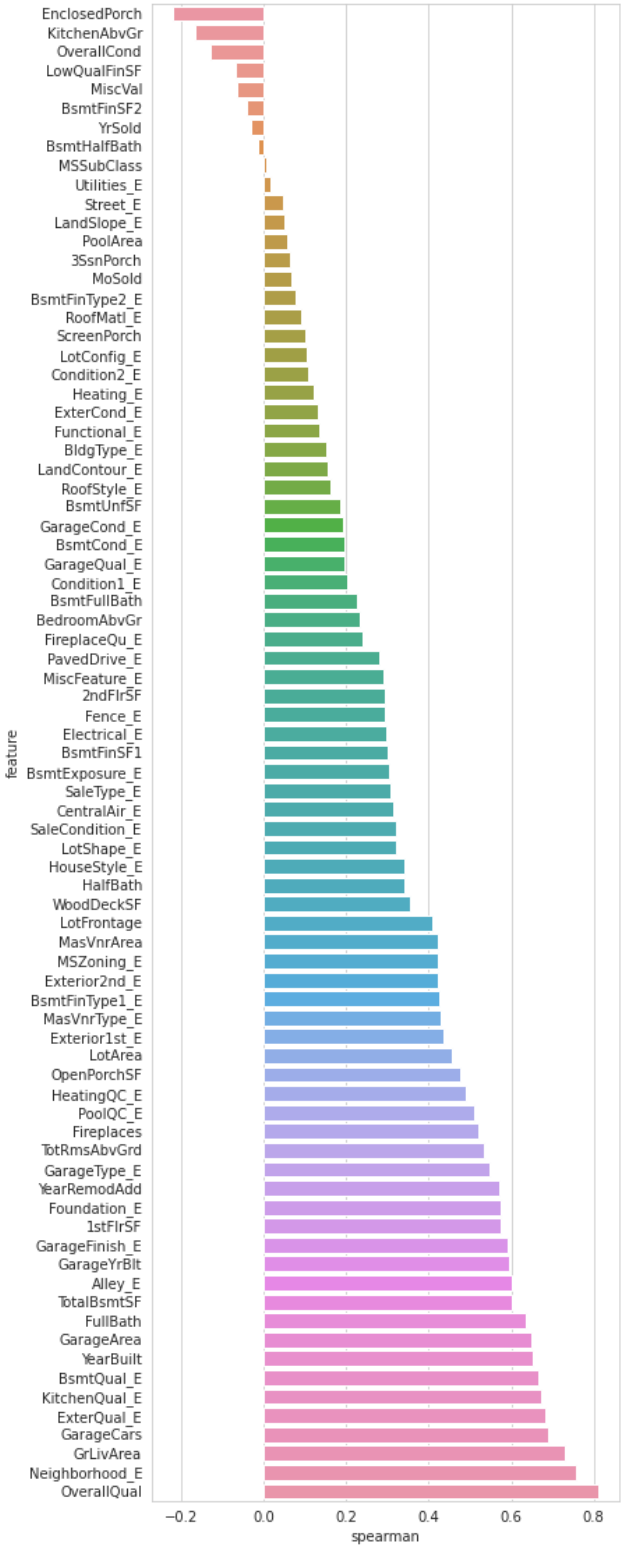


Top 8 skewed features :

Numerical Features	**Skewness**
PoolQC_E	22.954840
PoolArea	15.119426
3SsnPorch	8.924822
LowQualFinSF	8.744143
MiscFeature_E	7.305522
Alley_E	5.895238
MiscVal	5.597060
BsmtHalfBath	3.788243

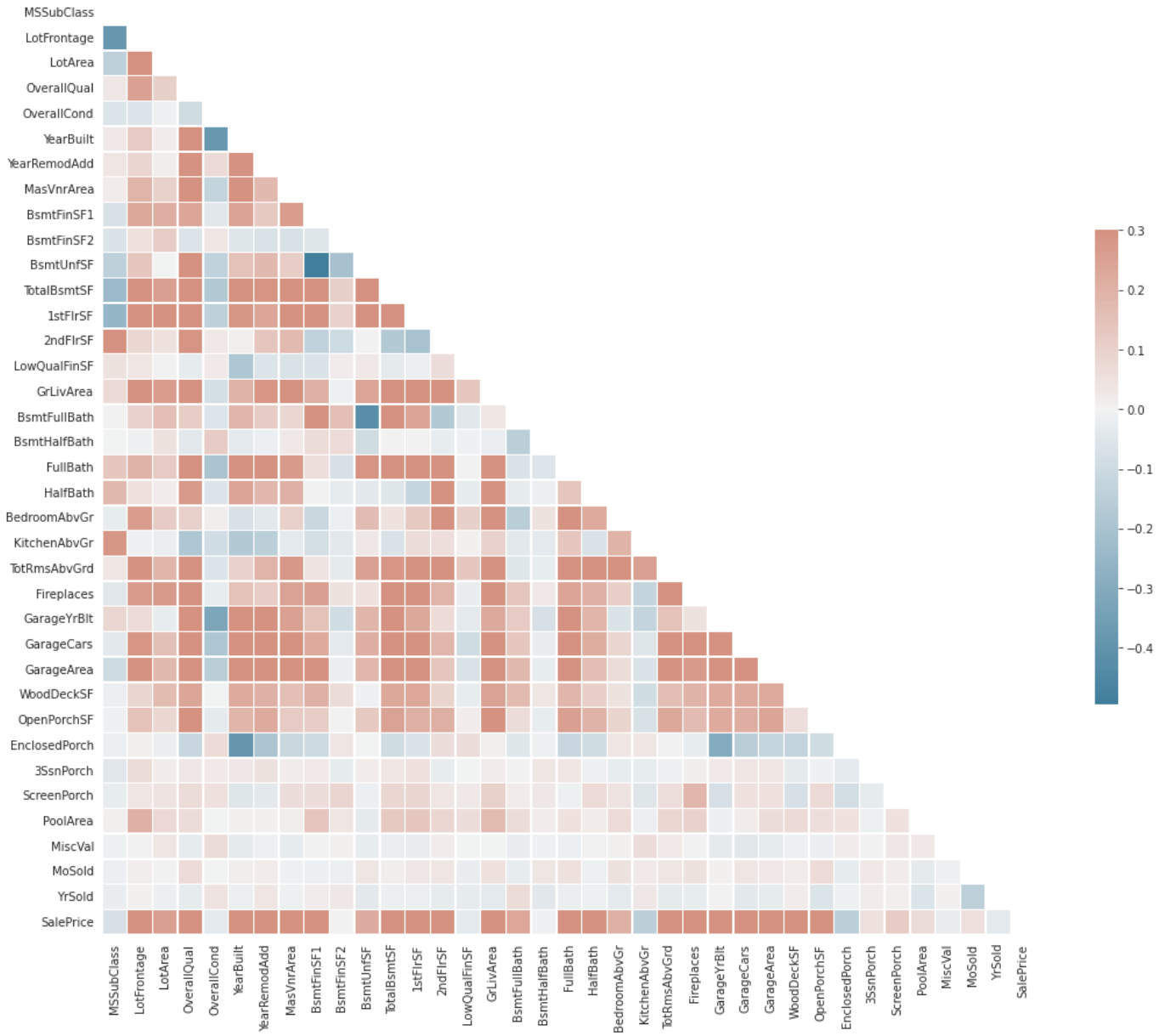
Correlation

After performing Spearman rank-order correlation on categorical variables, 43 new encoded features are added, correlation plot -

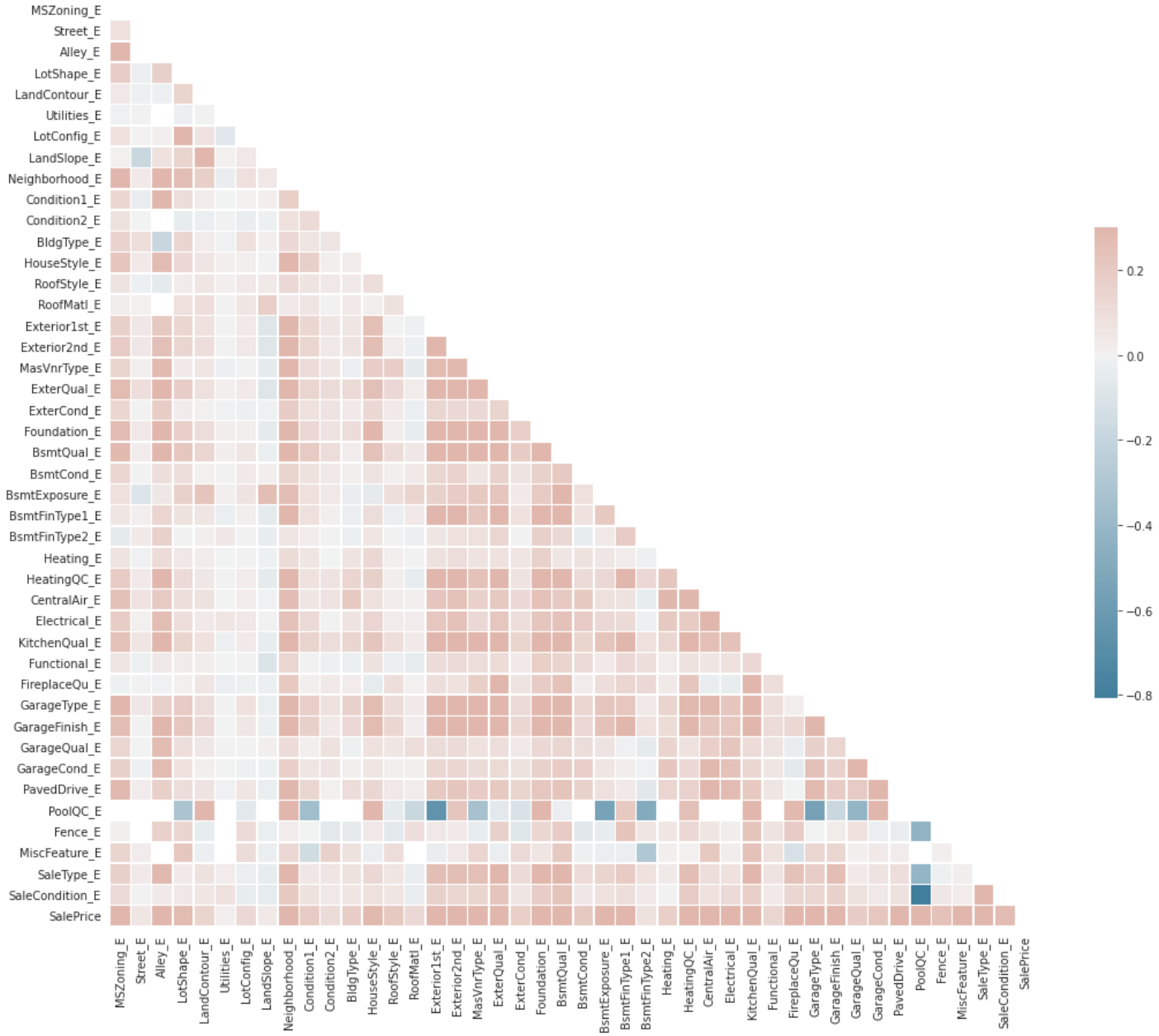


Heatmaps

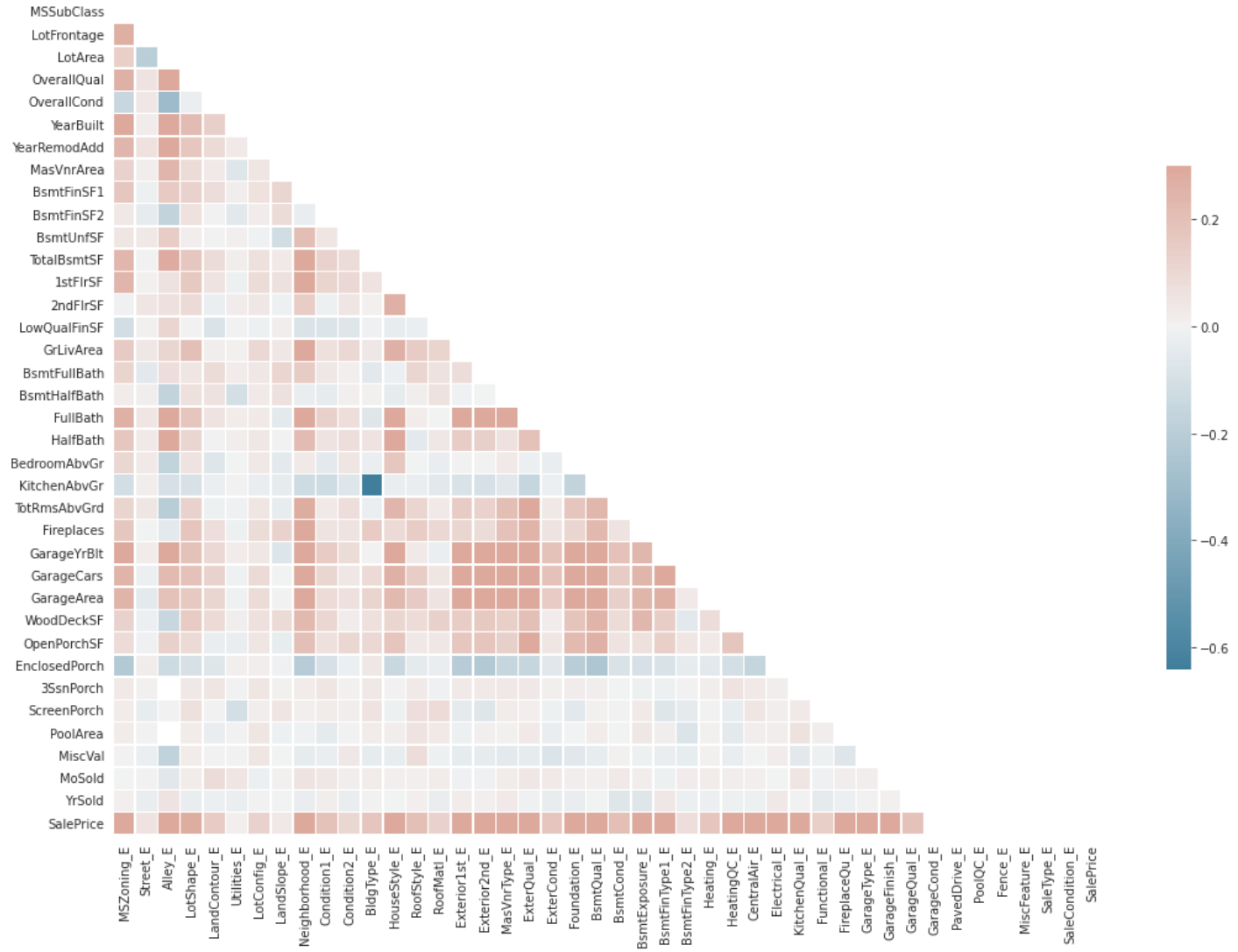
Correlation between **quantitative features and Sale Price -**


MSSubClass
LotFrontage
LotArea
OverallQual
OverallCond
YearBuilt
YearRemodAdd
MasVnrArea
BsmtFinSF1
BsmtFinSF2
BsmtUnfSF
TotalBsmtSF
1stFlrSF
2ndFlrSF
LowQualFinSF
GrLivArea
BsmtFullBath
BsmtHalfBath
FullBath
HalfBath
BedroomAbvGr
KitchenAbvGr
TotRmsAbvGrd
Fireplaces
GarageYrBlt
GarageCars
GarageArea
WoodDeckSF
OpenPorchSF
EnclosedPorch
3SsnPorch
ScreenPorch
PoolArea
MiscVal
MoSold
YrSold
SalePrice
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.4

Correlation between **qualitative features (encoded) and Sale Price –**



Correlation between **quantitative and qualitative features (encoded) -**



Data Preparation

- Log-transformed SalePrice target column
- Converted to string datatype : 'MoSold', 'YrSold', 'MSSubClass'
- Drop “Utilities” feature since majority of its value (99%) lies in the train set
- Filled missing values with most-frequent category for ‘Electrical’, ‘Exterior1st’, ‘SaleType’, ‘Exterior2nd’, ‘LotFrontage’, ‘MSZoning’ and ‘GarageYrBlt’
- Filled missing value with “NA” or “None” for rest of the categorical features according to the data documentation.
- Filled missing values with 0 for numerical features
- Applied one hot encoding for categorical features, ended with 373 features

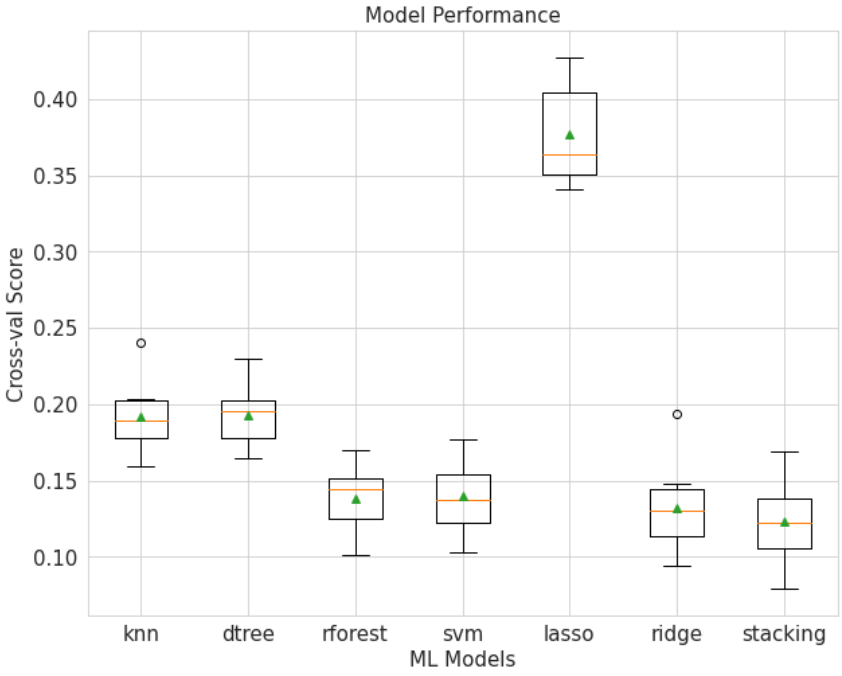
Modeling

Classic Machine Learning Models

Base model score after applying 10-fold cross validation on preprocessed train dataset -

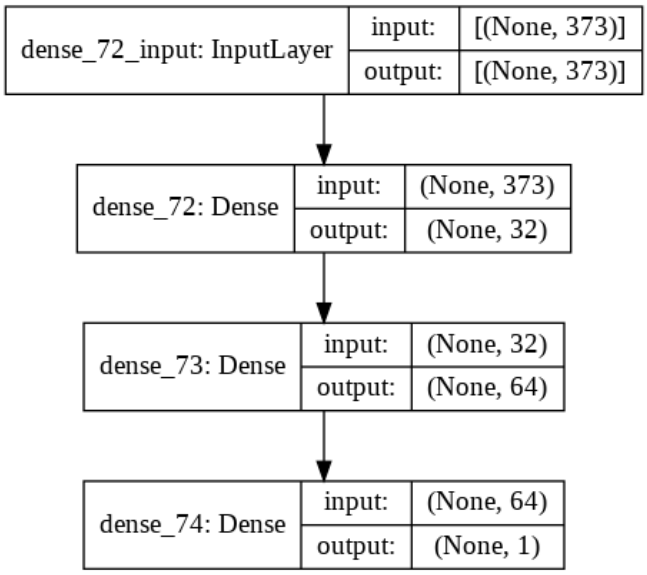
Model	RMSE Score (Mean)	RMSE Score (std)
Ridge Regression	0.1256	0.0244
LASSO Regression	0.1231	0.0251
Elastic net	0.1230	0.0251
Random Forest	0.1392	0.0220
XGBRegressor	0.1253	0.0188

Stacking Ensemble Model performance with base models

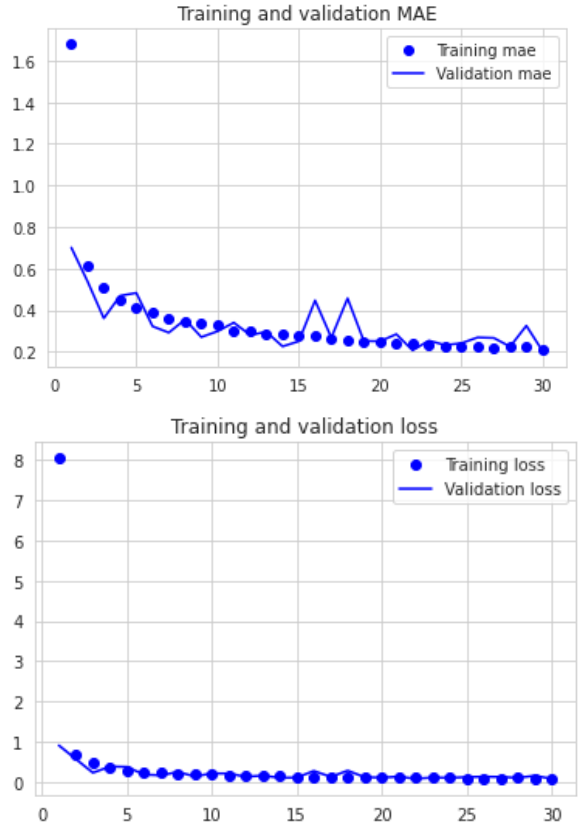


Deep learning Model

Architecture -



Performance (30 epochs, validation MAE ~ 0.2) -



Comparatively Deep Learning model performed worse, more optimization is needed, i.e., dropout, number of neuron or layers, tuning learning rate etc.

Future Work

- More hyperparameter optimization is needed
- Feature selections techniques such as – mutual information, chi-square will be explored