Project summary

The purpose of this project is to build machine learning models such as random forest, gradient boosting and stack models of the two to predict the sale prices of a house. This project is from the kaggle competition. In this summary I will outline four things:

1. Data source and data description
2. Initial data exploratory
3. Steps to clean data and select features
4. Build random forest, gradient boosting and stack model

**I. Data source and data description**

The project is from kaggle competition here is the project link include the data sources and the description <https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data>.

We have the train.csv file used to train the model, test.csv file used to test the model.

Data features description:

* 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
* MoSold: Month Sold
* YrSold: Year Sold
* SaleType: Type of sale
* SaleCondition: Condition of sale

**II. Initial data exploratory**

The code for the data exploratory are in data\_exploratory.R

Summary of the exploratory data result:

The dataset contains 1460 observation and 81 variables. Sale Price is the response variable.Id variables are used to keep count, so it not useful here. The other 79 predictors variable seem like can be used to build the model. And we notice something that for some of the categorical features attribute such as:

"Alley","BsmtQual","BsmtCond","BsmtExposure","BsmtFinType1","BsmtFinType2",

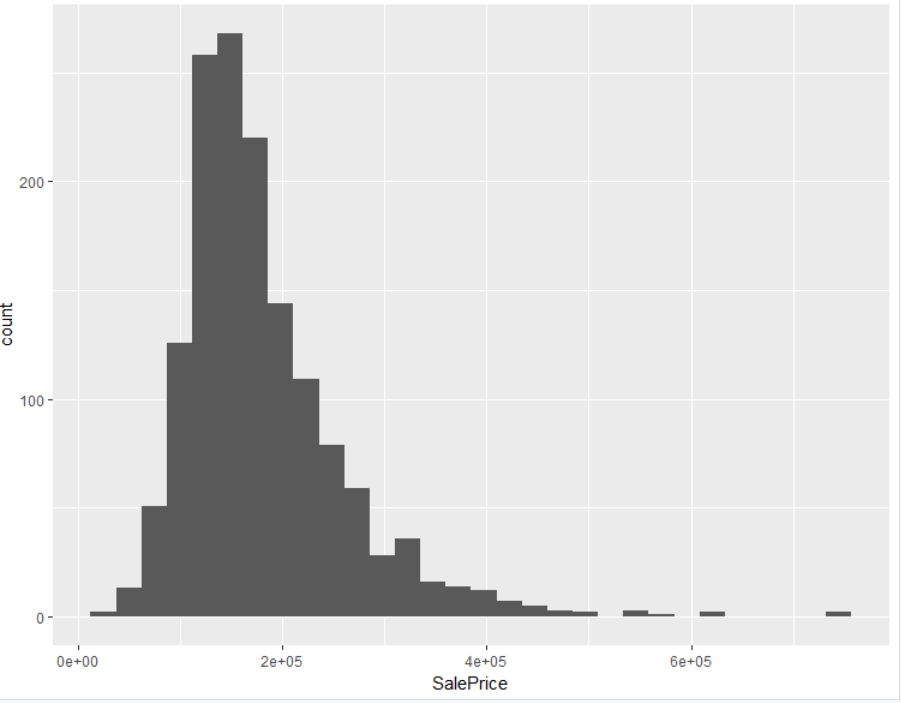
"FireplaceQu","GarageType","GarageFinish","GarageQual","GarageCond", "PoolQC","Fence","MiscFeature"

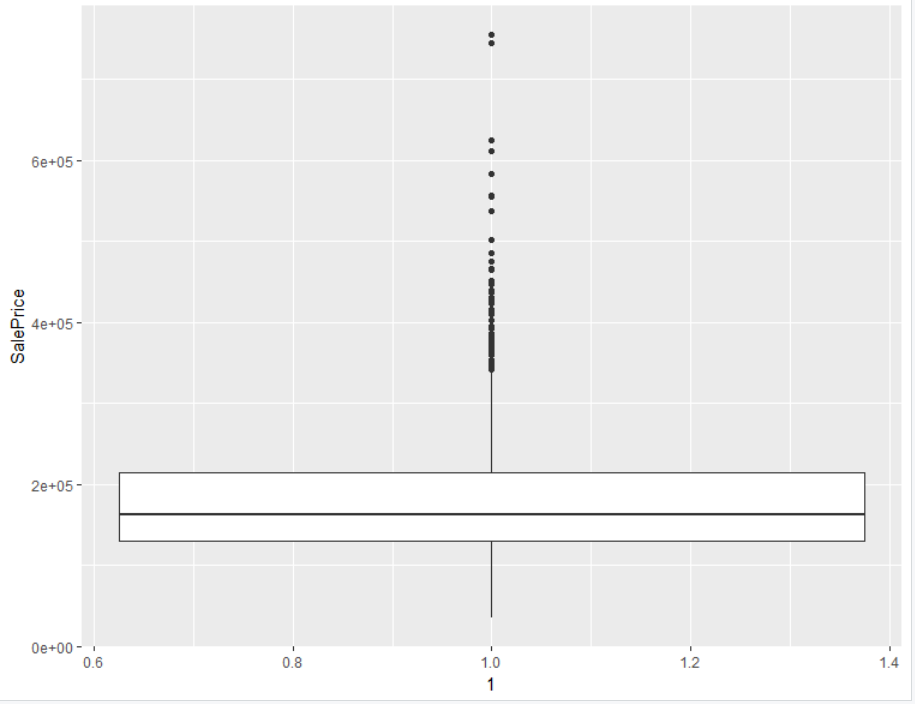
Contain NA but this does not mean missing value, but it means another factor levels

We will need to rename this NA factor level to None factor level so our process of building the model can use this information. There are some other features such as:

"MSSubClass", "MoSold"," YrSold"

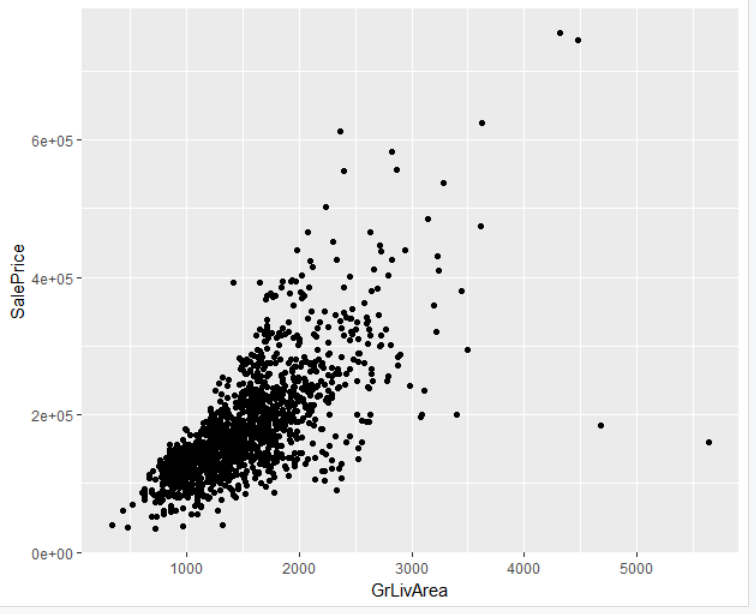
That is code as numeric type in the data, but we think they are more like categorical features so we will convert this. We also plot the Sale Price response variable to see what characteristic it have





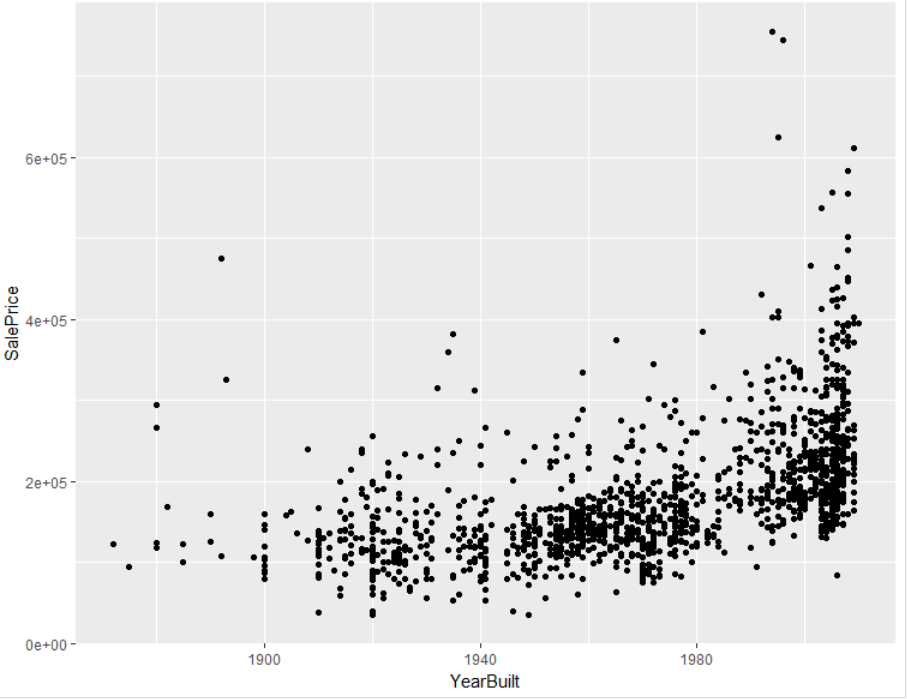
Seem like most of the house is in the 100k-250k range. The history graph is right skewed this indicate we might need to use log transformation on the data to get normally distributed data, and the box plot shows some houses are expensive this maybe outlier that we need to watch for when building the model.

Some exploratory variables are especially interested like the "GrLiveArea" "YearBuilt", "OverallCond" and "OveralQual" of the house. So we plot those

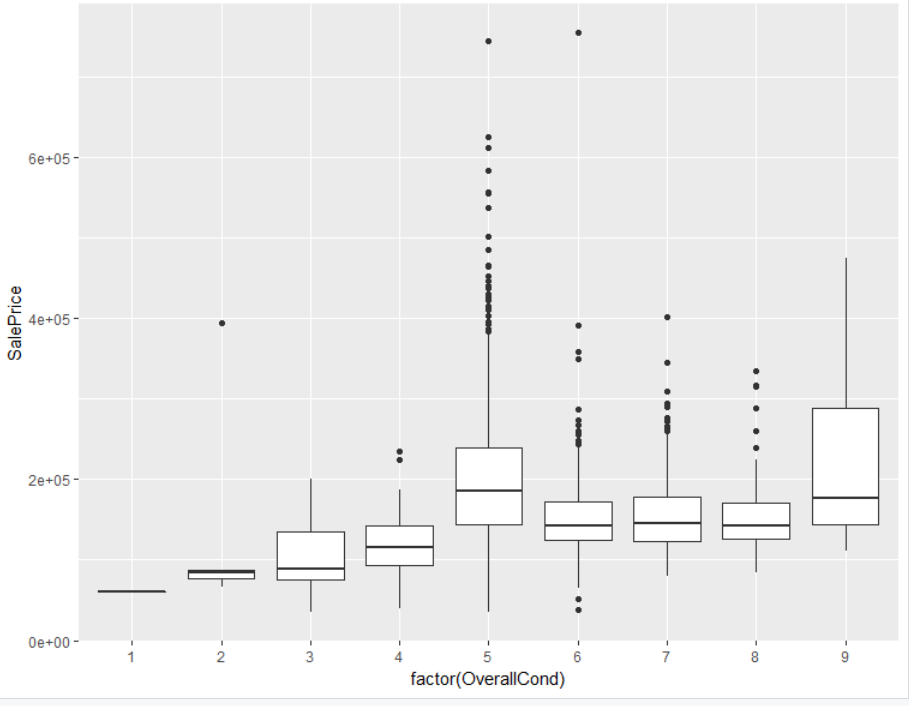


There are two outliers here with "GrLivArea" greater than 4000, but the price is less than 200000

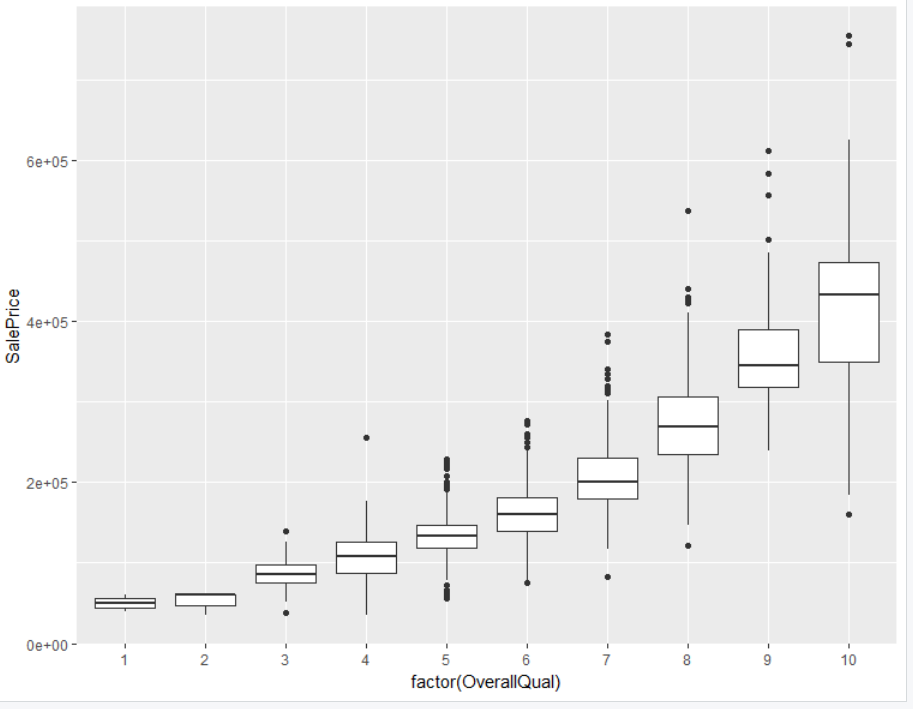
We will remove this outliers later.



Seem like we have a positive relationship between year built and sale price here. this can be used as a good predictor



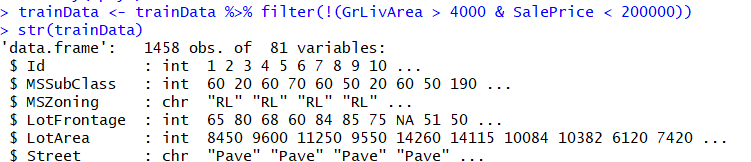
The sale price tend to increase as the overall condition rating go from 1 to 5 but does not change much or might even decrease as the over condition rating go from 5 to 9.



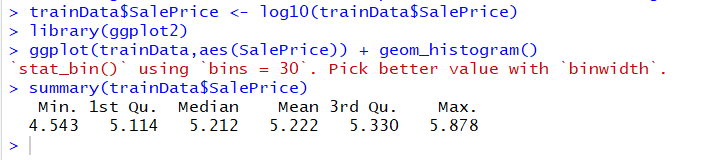
We see strong positive relationship here between SalePrice and overall quality rating. Seem to be a good predictor.

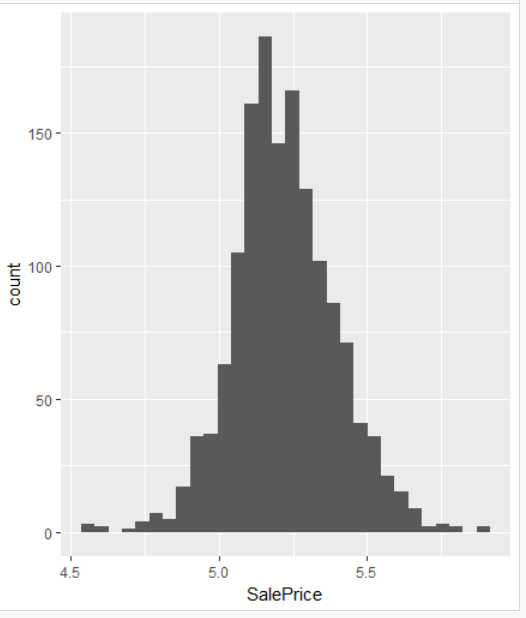
III. Steps to clean data and select features

1. Remove outliers

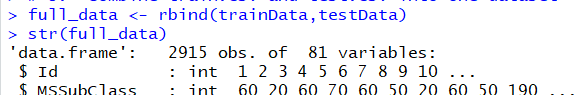


1. Log transform "SalePrice" response variable to remove right skew

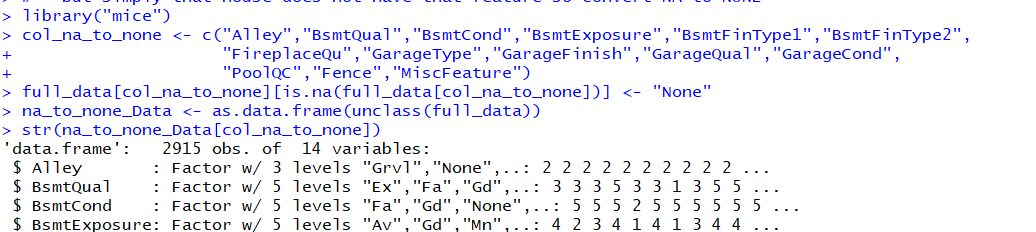




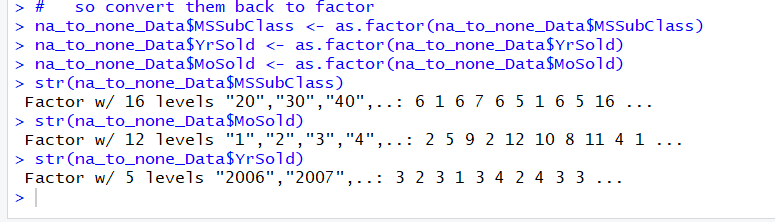
1. Combine train.csv and test.csv into one dataset



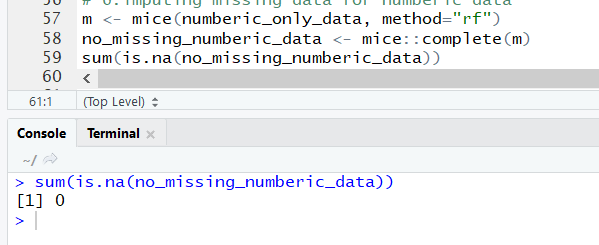
1. Convert Na to None for those columns that have NA as another factor level but not missing data



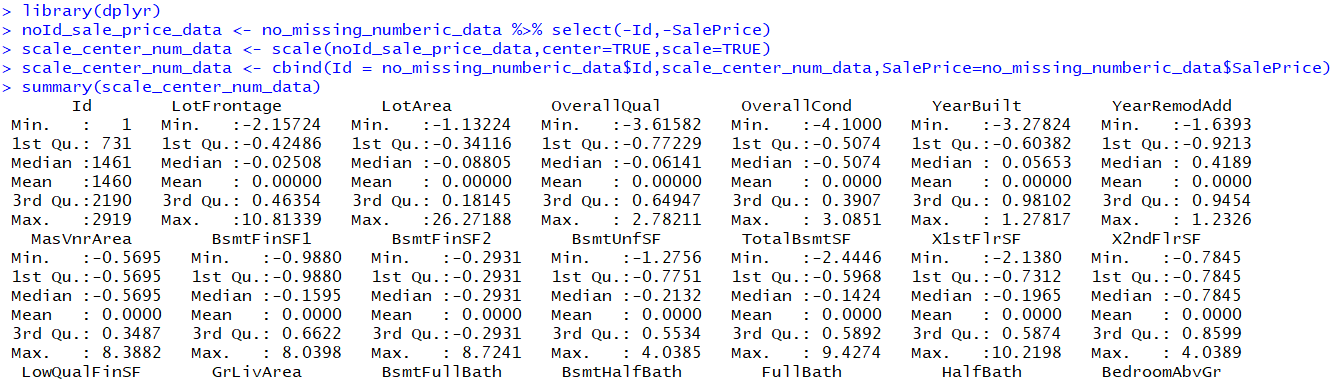
1. Convert to categorical features that are coded as numeric features



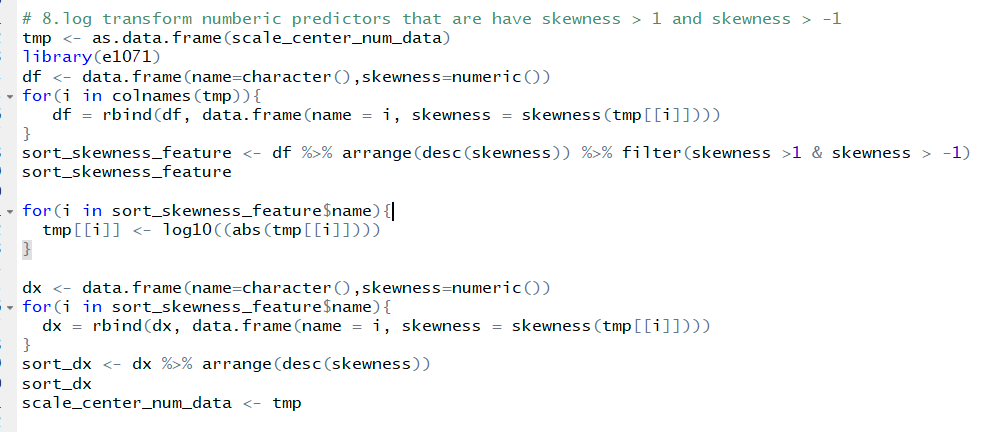
1. Impute numeric data using method rf in mice package



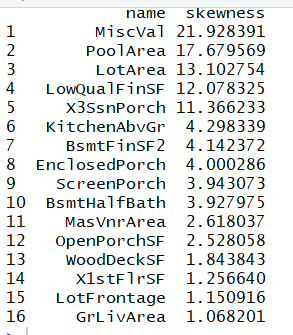
1. Scale and center numeric features except "ID" and "SalePrice" features



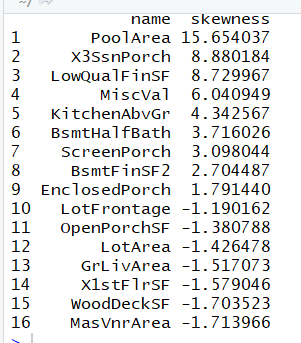
1. log transform numeric predictors to remove skew



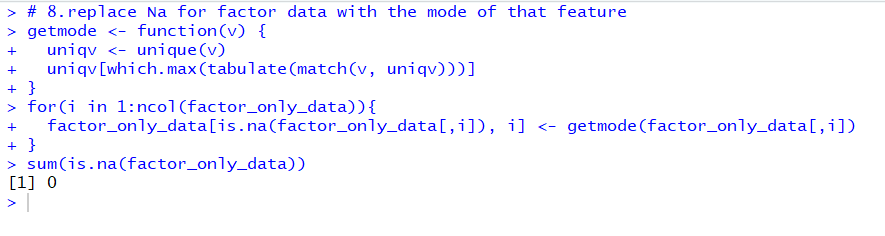
Before log10 transform



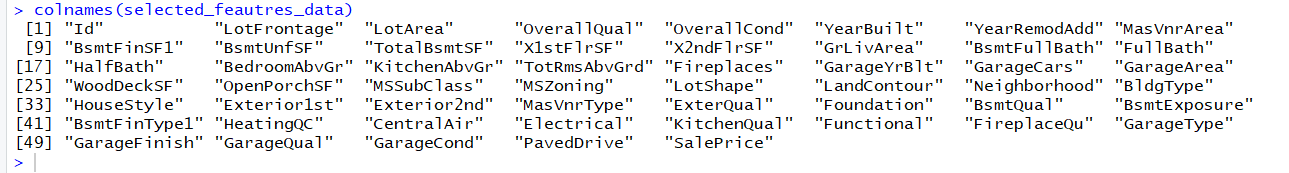
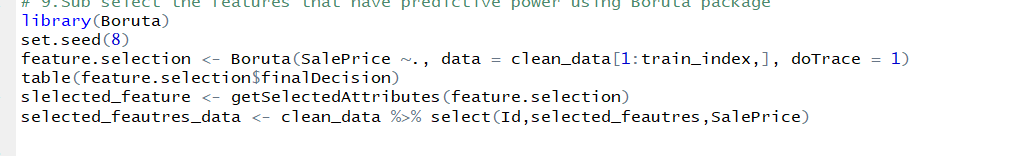
After log10 transform



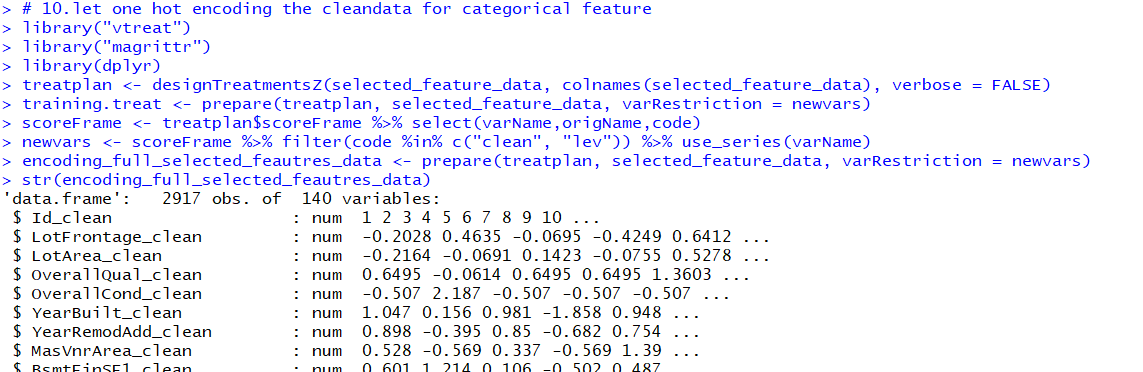
1. For categorical feature replace missing value with the mode of that features

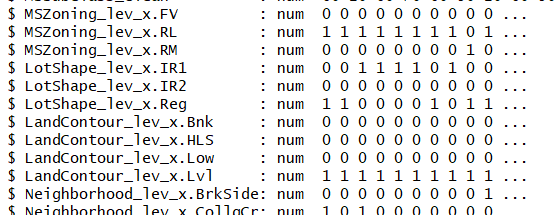


1. Subselect the features that have predictive power using Boruta package



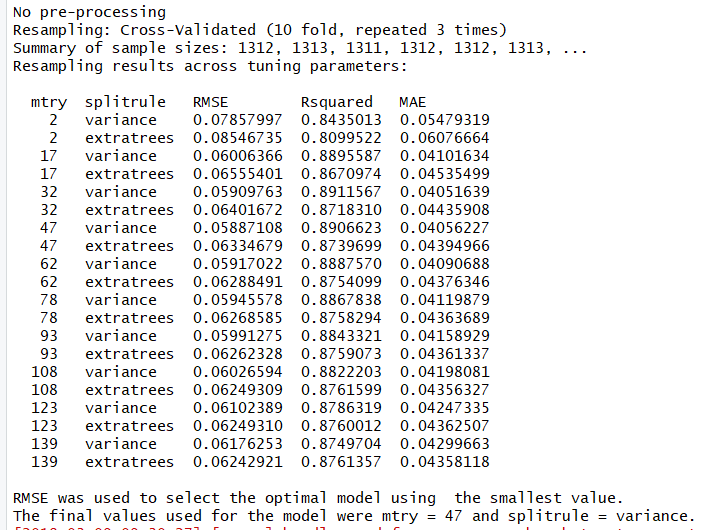
1. One hot encoding the categorical features.



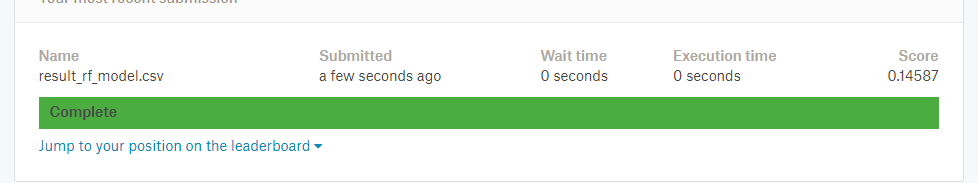


1. Build random forest, gradient boosting and stack model
2. Random forest model

After data have been cleaned in step III we build the model using 10-fold cross-validation repeated three times with tuneLenghth equal 10

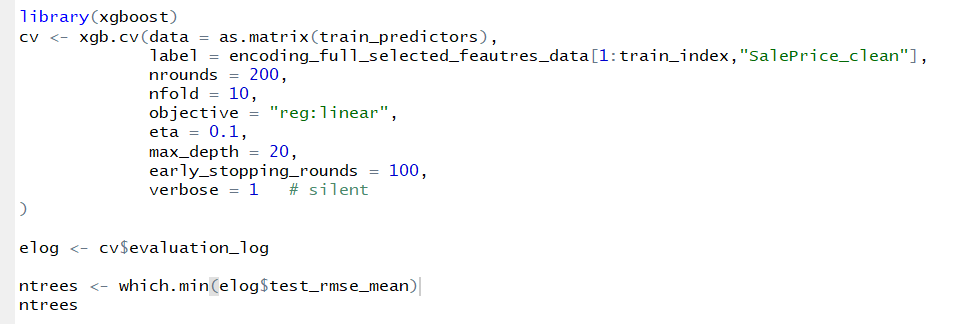


Predict test data result



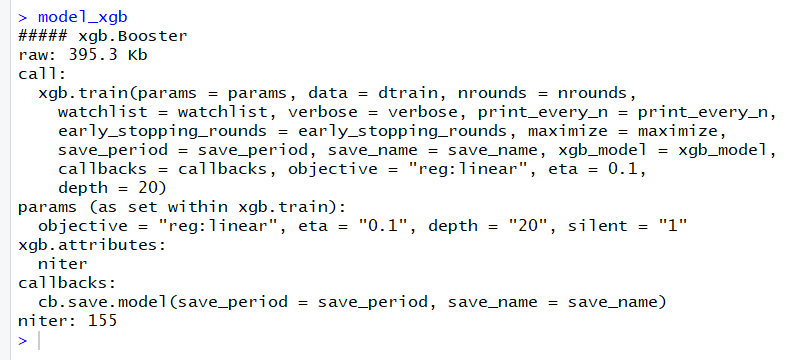
1. Gradient Boosting model

Finding the optimal number of the tree to use to build the gradient boosting model

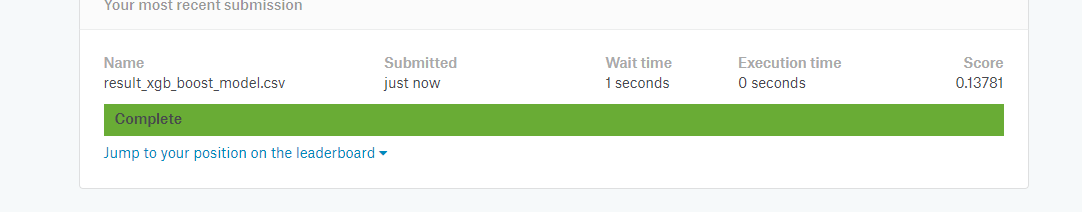




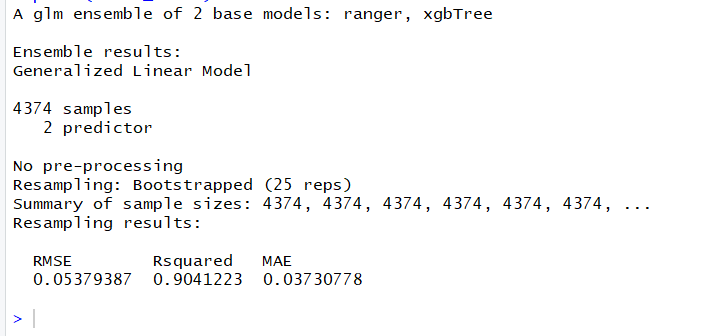
Building the model using ntrees



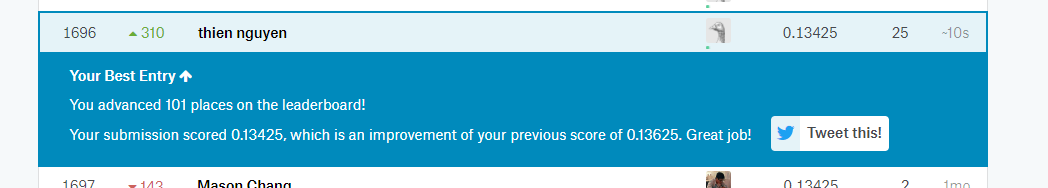
Predict test data result



1. Stack both models



Predict test data result



References

<https://www.kaggle.com/c/house-prices-advanced-regression-techniques>

Forte, R. M. (2015). *Mastering predictive analytics with R*. Packt Publishing Ltd.

Lesmeister, C. (2017). *Mastering machine learning with r*. Packt Publishing Ltd.

<https://www.datacamp.com/courses/machine-learning-toolbox>

https://machinelearningmastery.com/machine-learning-ensembles-with-r/