CASE STUDY ON HOUSING DATA

- By Sanjay VS

OBJECTIVE

The objective of this case study is to analyze the given house sales data and to selectively identify and filter the factors that affect sales price of a house and predict the sales price of the test data given.

DATA PRE-PROCESSING

There are 81 variables in this dataset. The data variables and their descriptions are given in the file data_descriptions. We have 1460 total observations in the given dataset.

Checking for NA's :

When we look at the data we can see that there are a lot of variables with missing data in it. A data variable having an NA can mean one of two things i.e the value was randomly found missing or that it could be because the particular feature is not there in the given property. We find that the following are the variables that have missing values :

- PoolQC: Pool quality
- MiscFeature: Miscellaneous feature not covered in other categories
- Alley: Type of alley access to property
- Fence: Fence qualityFireplaceQu: Fireplace quality
- FireplaceQu: Fireplace quality
- LotFrontage: Linear feet of street connected to property
- GarageCond: Garage condition
- GarageQual: Garage quality
- GarageFinish: Interior finish of the garage

- GarageYrBlt: Year garage was built
- GarageType: Garage location
- BsmtExposure: Refers to walkout or garden level walls
- BsmtCond: Evaluates the general condition of the basement
- BsmtQual: Evaluates the height of the basement
- BsmtFinType2: Rating of basement finished area (if multiple types)
- BsmtFinType1: Rating of basement finished area
- BsmtFinSF1: Type 1 finished square feet
- BsmtFinSF2: Type 2 finished square feet
- BsmtUnfSF: Unfinished square feet of basement area
- BsmtFullBath: Basement full bathrooms
- BsmtHalfBath: Basement half bathrooms
- MasVnrType: Masonry veneer type
- MasVnrArea: Masonry veneer area in square feet
- MSZoning: Identifies the general zoning classification of the sale
- Utilities: Type of utilities available
- Functional: Home functionality
- SaleType: Type of sale
- KitchenQual: Kitchen quality
- Electrical: Electrical system
- Exterior1st
- Exterior2nd

MISSING DATA IMPUTATION

Since we have a lot of variables that have missing data we realize the need of carefully trying to impute these missing observations with meaningful values. Each variable has its own set of properties so we might need to use different methods for imputation. The following is how I handled imputation for different variables :

Pool Data (PoolQC)

We have the levels Excellent(Ex), Good(Gd), Average(TA), Fair(Fa) and NA(No Pool) in this variable. The distribution for these is as following :

```
##   Ex   Fa   Gd   NA's
##    4    2    4   2909
```

We also find that there are 3 pools where the pool area is greater than 0 but the pool quality is not mentioned. Rest of the rows with PoolQc as NA has PoolArea = 0. So we can reassign None for "No pool" and make it another factor instead of NA's. The other 3 PoolArea values are given Good Quality based on majority.

MiscFeature, Alley, Fence, FireplaceQu

As per the data description NA in MiscFeature means None. Therefore we can directly replace the misleading NA by None.

Similarly for Type of Alley NA means No alley access, for Fence Quality NA means No fence and for Fireplace Quality NA means No Fireplace. So we can replace the missing values in these variables directly by None.

Garage

There are 157 observations test and train included with GarageType = NA, 1 observation with both GarageArea and GarageCars as NA and 159 observations with GarageYrBlt, GarageFinish, GarageQual and GarageCond as Nas.

We can directly replace the subsequent Garage columns with "None" when GarageArea =0. We predict the value of GarageYrBlt using rpart and the parameters GarageType, GarageFinish, GarageQual, GarageCond, GarageCars,GarageArea and YearBuilt.

The remaining Garage values missing are of a garage built in 1950. We find that in the year 1950 all the Garages have GarageQual as Unf, GarageFinish and GarageCond as TA. So we replace it in our missing observation.

MSZoning

We find out the values for general zoning using the parameters neighbourhood data and various conditions.(Condition1 and Condition2)

Basement

We find that in total there are 79 rows with NA for basement data. There is one row with all basement data missing so we assume TotalBsmtSF to be 0.

We use rpart to predict the Basement Condition for the missing observations. The parameters we use for prediction are BsmtExposure, BsmtCond, BsmtQual, BsmtFinType1, BsmtFinType2, TotalBsmtSF, YearBuilt.

Next we put missing BsmtFullBath and BsmtHalfBath as 0.

Utilities

We can see that all the rows have Utilities set to AllPub so we can just replace the missing value with AllPub.

Kitchen Quality

The parameters used in rpart to predict the values for Kitchen Qualiy are BldgType, HouseStyle, OverallQual, OverallCond, YearBuilt, YearRemodAdd.

Functional

The data description asks us to assume it is Typical unless specified so we are just going to convert missing Functional values to Typ

Electrical

The parameters used for predicting missing values for Electrical are BldgType, HouseStyle, OverallQual, OverallCond, YearBuilt, YearRemodAdd, Electrical

Exterior

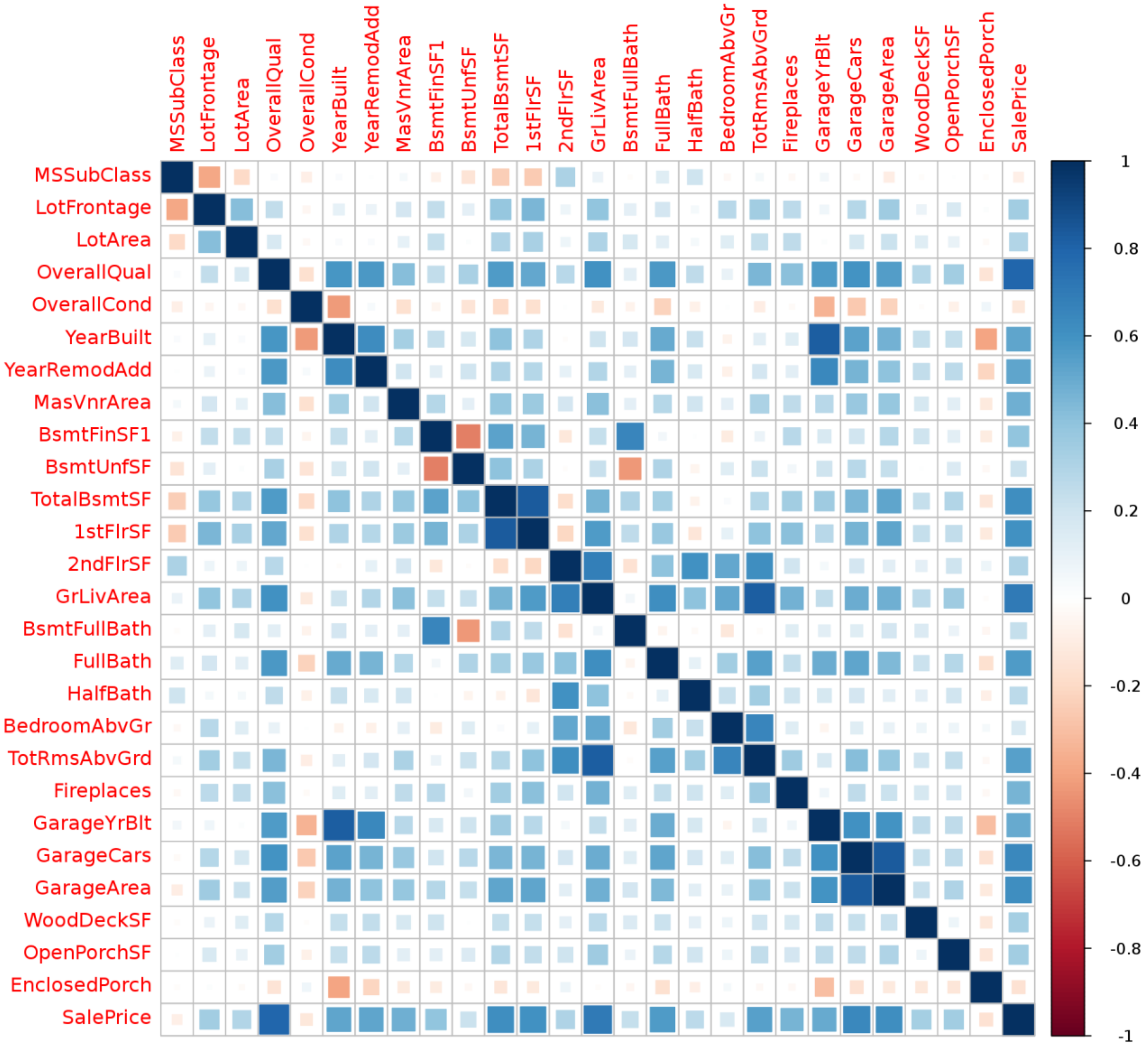
The parameters used to predict the missing values for Exterior1st and Exterior2nd variables are BldgType, HouseStyle, OverallQual, OverallCond, YearBuilt, YearRemodAdd, RoofStyle, RoofMatl, Exterior1st, Exterior2nd, "MasVnrType", MasVnrArea, ExterQual, ExterCond.

SaleType

The missing values in SaleType are just replaced by the “other” factor already given in the data description.

CORRELATION PLOT

Following is the correlation plot showing important features.



RANDOM FOREST

The classification model I have used to predict the SalePrice is random forest.

The problem statement asks us to build a model which would predict the Sales price of each house.

By using the importance factor in random forest predictor we come to know that the most important factors in predicting SalesPrice are OverallQual, Neighbourhood, GrLivArea, ExterQual, GarageCars, KitchenQual.

There are 98 factors in GarageYrBlt so we need to handle this as random forest cannot handle any factor variable with more than 53 categories. So we convert this variable into numeric.

Also we need to convert all the character variables to Factors. To do that we use the following lines of code :

```
train[sapply(train, is.character)] <- lapply(train[sapply(train, is.character)], as.factor)

test[sapply(test, is.character)] <- lapply(test[sapply(test, is.character)], as.factor)
```

Next we convert the SalesPrice to log values to take care of skewness. Now we build a random forest model over the test data and predict sales price and convert it back to normal values from logarithmic values.

CONCLUSION

What we find is that the main factors affecting the SalesPrice are of the houses are OverallQual, Neighbourhood, GrLivArea, ExterQual.

A model with decent accuracy can be built with the amount of data given. Another approach where we consider the majority output of Random forest, knn and naïve bayes is also worth a try in such a scenario but it kind of gives similar results.