KagglePrediction

Mohanraj

July 31, 2019

This Document explains how to predict price with many variables in place Load the Training data

```
setwd("D:\\Kaggle\\PredictHouse")
housing<-read.csv("train.csv")
df_housing=housing
str(df_housing)
```

```
## 'data.frame':    1460 obs. of  81 variables:
##  $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ MSSubClass   : int  60 20 60 70 60 50 20 60 50 190 ...
##  $ MSZoning     : Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5
4 ...
##  $ LotFrontage  : int  65 80 68 60 84 85 75 NA 51 50 ...
##  $ LotArea      : int  8450 9600 11250 9550 14260 14115 10084 10382 6120
7420 ...
##  $ Street       : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2
...
##  $ Alley        : Factor w/ 2 levels "Grvl","Pave": NA NA NA NA NA NA NA
NA NA NA ...
##  $ LotShape     : Factor w/ 4 levels "IR1","IR2","IR3",..: 4 4 1 1 1 1 4 1
4 4 ...
##  $ LandContour  : Factor w/ 4 levels "Bnk","HLS","Low",..: 4 4 4 4 4 4 4 4
4 4 ...
##  $ Utilities    : Factor w/ 2 levels "AllPub","NoSeWa": 1 1 1 1 1 1 1 1 1
1 ...
##  $ LotConfig    : Factor w/ 5 levels "Corner","CulDSac",..: 5 3 5 1 3 5 5
1 5 1 ...
##  $ LandSlope    : Factor w/ 3 levels "Gtl","Mod","Sev": 1 1 1 1 1 1 1 1 1
1 ...
##  $ Neighborhood : Factor w/ 25 levels "Blmngtn","Blueste",..: 6 25 6 7 14
12 21 17 18 4 ...
##  $ Condition1   : Factor w/ 9 levels "Artery","Feedr",..: 3 2 3 3 3 3 3 5
1 1 ...
##  $ Condition2   : Factor w/ 8 levels "Artery","Feedr",..: 3 3 3 3 3 3 3 3
3 1 ...
##  $ BldgType     : Factor w/ 5 levels "1Fam","2fmCon",..: 1 1 1 1 1 1 1 1 1
2 ...
##  $ HouseStyle   : Factor w/ 8 levels "1.5Fin","1.5Unf",..: 6 3 6 6 6 1 3 6
1 2 ...
##  $ OverallQual  : int  7 6 7 7 8 5 8 7 7 5 ...
##  $ OverallCond  : int  5 8 5 5 5 5 5 6 5 6 ...
```

```
##  $ YearBuilt    : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939
...
##  $ YearRemodAdd : int  2003 1976 2002 1970 2000 1995 2005 1973 1950 1950
...
##  $ RoofStyle    : Factor w/ 6 levels "Flat","Gable",..: 2 2 2 2 2 2 2 2 2
2 ...
##  $ RoofMatl     : Factor w/ 8 levels "ClyTile","CompShg",..: 2 2 2 2 2 2 2
2 2 2 ...
##  $ Exterior1st  : Factor w/ 15 levels "AsbShng","AsphShn",..: 13 9 13 14
13 13 13 7 4 9 ...
##  $ Exterior2nd  : Factor w/ 16 levels "AsbShng","AsphShn",..: 14 9 14 16
14 14 14 7 16 9 ...
##  $ MasVnrType   : Factor w/ 4 levels "BrkCmn","BrkFace",..: 2 3 2 3 2 3 4
4 3 3 ...
##  $ MasVnrArea   : int  196 0 162 0 350 0 186 240 0 0 ...
##  $ ExterQual    : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 4 3 4 3 4 4
4 ...
##  $ ExterCond    : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5
5 ...
##  $ Foundation   : Factor w/ 6 levels "BrkTil","CBlock",..: 3 2 3 1 3 6 3 2
1 1 ...
##  $ BsmtQual     : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 3 3 4 3 3 1 3 4
4 ...
##  $ BsmtCond     : Factor w/ 4 levels "Fa","Gd","Po",..: 4 4 4 2 4 4 4 4 4
4 ...
##  $ BsmtExposure : Factor w/ 4 levels "Av","Gd","Mn",..: 4 2 3 4 1 4 1 3 4
4 ...
##  $ BsmtFinType1 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 3 1 3 1 3 3 3 1
6 3 ...
##  $ BsmtFinSF1   : int  706 978 486 216 655 732 1369 859 0 851 ...
##  $ BsmtFinType2 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 6 6 6 6 6 6 6 2
6 6 ...
##  $ BsmtFinSF2   : int  0 0 0 0 0 0 0 32 0 0 ...
##  $ BsmtUnfSF    : int  150 284 434 540 490 64 317 216 952 140 ...
##  $ TotalBsmtSF  : int  856 1262 920 756 1145 796 1686 1107 952 991 ...
##  $ Heating      : Factor w/ 6 levels "Floor","GasA",..: 2 2 2 2 2 2 2 2 2
2 ...
##  $ HeatingQC    : Factor w/ 5 levels "Ex","Fa","Gd",..: 1 1 1 3 1 1 1 1 3
1 ...
##  $ CentralAir   : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Electrical   : Factor w/ 5 levels "FuseA","FuseF",..: 5 5 5 5 5 5 5 5 2
5 ...
##  $ X1stFlrSF    : int  856 1262 920 961 1145 796 1694 1107 1022 1077 ...
##  $ X2ndFlrSF    : int  854 0 866 756 1053 566 0 983 752 0 ...
##  $ LowQualFinSF : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ GrLivArea    : int  1710 1262 1786 1717 2198 1362 1694 2090 1774 1077
...
##  $ BsmtFullBath : int  1 0 1 1 1 1 1 1 0 1 ...
##  $ BsmtHalfBath : int  0 1 0 0 0 0 0 0 0 0 ...
##  $ FullBath     : int  2 2 2 1 2 1 2 2 2 1 ...
```

```
##  $ HalfBath     : int  1 0 1 0 1 1 0 1 0 0 ...
##  $ BedroomAbvGr : int  3 3 3 3 4 1 3 3 2 2 ...
##  $ KitchenAbvGr : int  1 1 1 1 1 1 1 1 2 2 ...
##  $ KitchenQual  : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 3 3 4 3 4 4
4 ...
##  $ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
##  $ Functional   : Factor w/ 7 levels "Maj1","Maj2",..: 7 7 7 7 7 7 7 7 3 7
...
##  $ Fireplaces   : int  0 1 1 1 1 0 1 2 2 2 ...
##  $ FireplaceQu  : Factor w/ 5 levels "Ex","Fa","Gd",..: NA 5 5 3 5 NA 3 5
5 5 ...
##  $ GarageType   : Factor w/ 6 levels "2Types","Attchd",..: 2 2 2 6 2 2 2 2
6 2 ...
##  $ GarageYrBlt  : int  2003 1976 2001 1998 2000 1993 2004 1973 1931 1939
...
##  $ GarageFinish : Factor w/ 3 levels "Fin","RFn","Unf": 2 2 2 3 2 3 2 2 3
2 ...
##  $ GarageCars   : int  2 2 2 3 3 2 2 2 2 1 ...
##  $ GarageArea   : int  548 460 608 642 836 480 636 484 468 205 ...
##  $ GarageQual   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 2
3 ...
##  $ GarageCond   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5
5 ...
##  $ PavedDrive   : Factor w/ 3 levels "N","P","Y": 3 3 3 3 3 3 3 3 3 3 ...
##  $ WoodDeckSF   : int  0 298 0 0 192 40 255 235 90 0 ...
##  $ OpenPorchSF  : int  61 0 42 35 84 30 57 204 0 4 ...
##  $ EnclosedPorch: int  0 0 0 272 0 0 0 228 205 0 ...
##  $ X3SsnPorch   : int  0 0 0 0 0 320 0 0 0 0 ...
##  $ ScreenPorch  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolArea     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ PoolQC       : Factor w/ 3 levels "Ex","Fa","Gd": NA NA NA NA NA NA NA
NA NA NA ...
##  $ Fence        : Factor w/ 4 levels "GdPrv","GdWo",..: NA NA NA NA NA 3
NA NA NA NA ...
##  $ MiscFeature  : Factor w/ 4 levels "Gar2","Othr",..: NA NA NA NA NA 3 NA
3 NA NA ...
##  $ MiscVal      : int  0 0 0 0 0 700 0 350 0 0 ...
##  $ MoSold       : int  2 5 9 2 12 10 8 11 4 1 ...
##  $ YrSold       : int  2008 2007 2008 2006 2008 2009 2007 2009 2008 2008
...
##  $ SaleType     : Factor w/ 9 levels "COD","Con","ConLD",..: 9 9 9 9 9 9 9
9 9 9 ...
##  $ SaleCondition: Factor w/ 6 levels "Abnorml","AdjLand",..: 5 5 5 1 5 5 5
5 1 5 ...
##  $ SalePrice    : int  208500 181500 223500 140000 250000 143000 307000
200000 129900 118000 ...
```

Above command helps to give the datatype and the factor variables of the data. Next Step is to find the missing values from the data provided

```
colSums(is.na(df_housing))
```

```
##            Id    MSSubClass      MSZoning   LotFrontage       LotArea
##             0             0             0           259             0
##        Street         Alley      LotShape   LandContour     Utilities
##             0          1369             0             0             0
##     LotConfig     LandSlope  Neighborhood    Condition1    Condition2
##             0             0             0             0             0
##      BldgType    HouseStyle   OverallQual   OverallCond     YearBuilt
##             0             0             0             0             0
##  YearRemodAdd     RoofStyle      RoofMatl   Exterior1st   Exterior2nd
##             0             0             0             0             0
##    MasVnrType    MasVnrArea     ExterQual     ExterCond    Foundation
##             8             8             0             0             0
##      BsmtQual      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1
##            37            37            38            37             0
##  BsmtFinType2    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating
##            38             0             0             0             0
##     HeatingQC    CentralAir    Electrical     X1stFlrSF     X2ndFlrSF
##             0             0             1             0             0
##  LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath
##             0             0             0             0             0
##      HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd
##             0             0             0             0             0
##    Functional    Fireplaces   FireplaceQu    GarageType   GarageYrBlt
##             0             0           690            81            81
##  GarageFinish    GarageCars    GarageArea    GarageQual    GarageCond
##            81             0             0            81            81
##    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch
##             0             0             0             0             0
##   ScreenPorch      PoolArea        PoolQC         Fence   MiscFeature
##             0             0          1453          1179          1406
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##             0             0             0             0             0
##     SalePrice
##             0
```

Let us remove fields which has more 50% missing values

```
df_housing=subset(df_housing,select = -
c(PoolQC,Fence,MiscFeature,Alley,FireplaceQu))
```

For fields LotFrontage & GarageYrBlt we are replacing the missing values with mean of the columns

```
mean_Lfrontage=mean(df_housing$LotFrontage,na.rm = TRUE)
df_housing$LotFrontage[is.na(df_housing$LotFrontage)]=mean_Lfrontage

mean_GarageYrBlt=mean(df_housing$GarageYrBlt,na.rm = TRUE)
df_housing$GarageYrBlt[is.na(df_housing$GarageYrBlt)]=mean_GarageYrBlt
```

```
mean_Lfrontage
```

```
## [1] 70.04996
```

```
mean_GarageYrBlt
```

```
## [1] 1978.506
```

Above method can be used to replace missing values or we can use libraries to impute missing values

```
library(missForest)
```

```
## Loading required package: randomForest
```

```
## randomForest 4.6-14
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
## Loading required package: foreach
```

```
## Loading required package: itertools
```

```
## Loading required package: iterators
```

```
df_housing.imp<-missForest(df_housing)
```

```
##   missForest iteration 1 in progress...done!
##   missForest iteration 2 in progress...done!
##   missForest iteration 3 in progress...done!
##   missForest iteration 4 in progress...done!
##   missForest iteration 5 in progress...done!
##   missForest iteration 6 in progress...done!
```

```
df_housing=df_housing.imp$ximp
df_housing<-na.omit(df_housing)
colSums(is.na(df_housing))
```

```
##            Id    MSSubClass      MSZoning   LotFrontage       LotArea
##             0             0             0             0             0
##        Street      LotShape   LandContour     Utilities     LotConfig
##             0             0             0             0             0
##     LandSlope  Neighborhood    Condition1    Condition2      BldgType
##             0             0             0             0             0
##    HouseStyle   OverallQual   OverallCond     YearBuilt  YearRemodAdd
##             0             0             0             0             0
##     RoofStyle      RoofMatl   Exterior1st   Exterior2nd    MasVnrType
##             0             0             0             0             0
##    MasVnrArea     ExterQual     ExterCond    Foundation      BsmtQual
##             0             0             0             0             0
##      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1  BsmtFinType2
##             0             0             0             0             0
```

```
##    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating     HeatingQC
##             0             0             0             0             0
##    CentralAir    Electrical     X1stFlrSF     X2ndFlrSF  LowQualFinSF
##             0             0             0             0             0
##     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath      HalfBath
##             0             0             0             0             0
##  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd    Functional
##             0             0             0             0             0
##    Fireplaces    GarageType   GarageYrBlt  GarageFinish    GarageCars
##             0             0             0             0             0
##    GarageArea    GarageQual    GarageCond    PavedDrive    WoodDeckSF
##             0             0             0             0             0
##   OpenPorchSF EnclosedPorch    X3SsnPorch   ScreenPorch      PoolArea
##             0             0             0             0             0
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##             0             0             0             0             0
##     SalePrice
##             0
```

Now we have dataframe without missing values and we can see all 0's Next things is to convert categorical value to numbers of level 2

```
levels(df_housing$Street)<-c(1,0)
df_housing$Street<- as.numeric(levels(df_housing$Street))[df_housing$Street]

levels(df_housing$Utilities)<-c(1,0)
df_housing$Utilities<-
as.numeric(levels(df_housing$Utilities))[df_housing$Utilities]

levels(df_housing$CentralAir)<-c(1,0)
df_housing$CentralAir<-
as.numeric(levels(df_housing$CentralAir))[df_housing$CentralAir]
```

Factor variable more than level 2 can be converted into dummy variables for the remaining categorical variables

```
dummy_MSZoning  <- data.frame(model.matrix( ~MSZoning , data = df_housing))
dummy_MSZoning  <- dummy_MSZoning[,-1]
dummy_LotShape<- data.frame(model.matrix( ~LotShape, data = df_housing))
dummy_LotShape<-dummy_LotShape[,-1]
dummy_LandContour<- data.frame(model.matrix( ~LandContour, data =
df_housing))
dummy_LandContour<-dummy_LandContour[,-1]
dummy_LotConfig<- data.frame(model.matrix( ~LotConfig, data = df_housing))
dummy_LotConfig<-dummy_LotConfig[,-1]
dummy_LandSlope<- data.frame(model.matrix( ~LandSlope, data = df_housing))
dummy_LandSlope<-dummy_LandSlope[,-1]
dummy_Neighborhood<- data.frame(model.matrix( ~Neighborhood, data =
df_housing))
dummy_Neighborhood<-dummy_Neighborhood[,-1]
```

```
dummy_Condition1<- data.frame(model.matrix( ~Condition1, data = df_housing))
dummy_Condition1<-dummy_Condition1[,-1]
dummy_Condition2<- data.frame(model.matrix( ~Condition2, data = df_housing))
dummy_Condition2<-dummy_Condition2[,-1]
dummy_BldgType<- data.frame(model.matrix( ~BldgType, data = df_housing))
dummy_BldgType<-dummy_BldgType[,-1]
dummy_HouseStyle<- data.frame(model.matrix( ~HouseStyle, data = df_housing))
dummy_HouseStyle<-dummy_HouseStyle[,-1]
dummy_RoofStyle<- data.frame(model.matrix( ~RoofStyle, data = df_housing))
dummy_RoofStyle<-dummy_RoofStyle[,-1]
dummy_RoofMatl<- data.frame(model.matrix( ~RoofMatl, data = df_housing))
dummy_RoofMatl<-dummy_RoofMatl[,-1]
dummy_Exterior1st<- data.frame(model.matrix( ~Exterior1st, data =
df_housing))
dummy_Exterior1st<-dummy_Exterior1st[,-1]
dummy_Exterior2nd<- data.frame(model.matrix( ~Exterior2nd, data =
df_housing))
dummy_Exterior2nd<-dummy_Exterior2nd[,-1]
dummy_MasVnrType<- data.frame(model.matrix( ~MasVnrType, data = df_housing))
dummy_MasVnrType<-dummy_MasVnrType[,-1]
dummy_ExterQual<- data.frame(model.matrix( ~ExterQual, data = df_housing))
dummy_ExterQual<-dummy_ExterQual[,-1]
dummy_ExterCond<- data.frame(model.matrix( ~ExterCond, data = df_housing))
dummy_ExterCond<-dummy_ExterCond[,-1]
dummy_Foundation<- data.frame(model.matrix( ~Foundation, data = df_housing))
dummy_Foundation<-dummy_Foundation[,-1]
dummy_BsmtQual<- data.frame(model.matrix( ~BsmtQual, data = df_housing))
dummy_BsmtQual<-dummy_BsmtQual[,-1]
dummy_BsmtCond<- data.frame(model.matrix( ~BsmtCond, data = df_housing))
dummy_BsmtCond<-dummy_BsmtCond[,-1]
dummy_BsmtFinType1<- data.frame(model.matrix( ~BsmtFinType1, data =
df_housing))
dummy_BsmtFinType1<-dummy_BsmtFinType1[,-1]
dummy_BsmtExposure<- data.frame(model.matrix( ~BsmtExposure, data =
df_housing))
dummy_BsmtExposure<-dummy_BsmtExposure[,-1]
dummy_BsmtFinType2<- data.frame(model.matrix( ~BsmtFinType2, data =
df_housing))
dummy_BsmtFinType2<-dummy_BsmtFinType2[,-1]
dummy_Heating<- data.frame(model.matrix( ~Heating, data = df_housing))
dummy_Heating<-dummy_Heating[,-1]
dummy_HeatingQC<- data.frame(model.matrix( ~HeatingQC, data = df_housing))
dummy_HeatingQC<-dummy_HeatingQC[,-1]
dummy_Electrical<- data.frame(model.matrix( ~Electrical, data = df_housing))
dummy_Electrical<-dummy_Electrical[,-1]
dummy_KitchenQual<- data.frame(model.matrix( ~KitchenQual, data =
df_housing))
dummy_KitchenQual<-dummy_KitchenQual[,-1]
dummy_Functional<- data.frame(model.matrix( ~Functional, data = df_housing))
dummy_Functional<-dummy_Functional[,-1]
```

```
dummy_GarageType<- data.frame(model.matrix( ~GarageType, data = df_housing))
dummy_GarageType<-dummy_GarageType[,-1]
dummy_GarageFinish<- data.frame(model.matrix( ~GarageFinish, data =
df_housing))
dummy_GarageFinish<-dummy_GarageFinish[,-1]
dummy_GarageQual<- data.frame(model.matrix( ~GarageQual, data = df_housing))
dummy_GarageQual<-dummy_GarageQual[,-1]
dummy_GarageCond<- data.frame(model.matrix( ~GarageCond, data = df_housing))
dummy_GarageCond<-dummy_GarageCond[,-1]
dummy_PavedDrive<- data.frame(model.matrix( ~PavedDrive, data = df_housing))
dummy_PavedDrive<-dummy_PavedDrive[,-1]
dummy_SaleType<- data.frame(model.matrix( ~SaleType, data = df_housing))
dummy_SaleType<-dummy_SaleType[,-1]
dummy_SaleCondition<- data.frame(model.matrix( ~SaleCondition, data =
df_housing))
dummy_SaleCondition<-dummy_SaleCondition[,-1]
```

Before adding the dummy variable to the original dataframe, It is good to remove the fields from the dataframe

```
df_housing=subset(df_housing,select=-
c(LotShape,LandContour,LotConfig,LandSlope,Neighborhood,Condition1,Condition2
,BldgType,HouseStyle,RoofStyle,RoofMatl,Exterior1st,Exterior2nd,MasVnrType,Ex
terQual,ExterCond,Foundation,BsmtQual,BsmtCond,BsmtFinType1,BsmtExposure,Bsmt
FinType2,Heating,HeatingQC,Electrical,KitchenQual,Functional,GarageType,Garag
eFinish,GarageQual,GarageCond,PavedDrive,SaleType,SaleCondition))
```

Add the dummy variables to the dataframe

```
df_housing_cat<- cbind(df_housing, dummy_MSZoning)
df_housing_cat<- cbind(df_housing_cat, dummy_LotShape)
df_housing_cat<- cbind(df_housing_cat, dummy_LandContour)
df_housing_cat<- cbind(df_housing_cat, dummy_LotConfig)
df_housing_cat<- cbind(df_housing_cat, dummy_LandSlope)
df_housing_cat<- cbind(df_housing_cat, dummy_Neighborhood)
df_housing_cat<- cbind(df_housing_cat, dummy_Condition1)
df_housing_cat<- cbind(df_housing_cat, dummy_Condition2)
df_housing_cat<- cbind(df_housing_cat, dummy_BldgType)
df_housing_cat<- cbind(df_housing_cat, dummy_HouseStyle)
df_housing_cat<- cbind(df_housing_cat, dummy_RoofStyle)
df_housing_cat<- cbind(df_housing_cat, dummy_RoofMatl)
df_housing_cat<- cbind(df_housing_cat, dummy_Exterior1st)
df_housing_cat<- cbind(df_housing_cat, dummy_Exterior2nd)
df_housing_cat<- cbind(df_housing_cat, dummy_MasVnrType)
df_housing_cat<- cbind(df_housing_cat, dummy_ExterQual)
df_housing_cat<- cbind(df_housing_cat, dummy_ExterCond)
df_housing_cat<- cbind(df_housing_cat, dummy_Foundation)
df_housing_cat<- cbind(df_housing_cat, dummy_BsmtQual)
df_housing_cat<- cbind(df_housing_cat, dummy_BsmtCond)
df_housing_cat<- cbind(df_housing_cat, dummy_BsmtFinType1)
df_housing_cat<- cbind(df_housing_cat, dummy_BsmtExposure)
```

```
df_housing_cat<- cbind(df_housing_cat, dummy_BsmtFinType2)
df_housing_cat<- cbind(df_housing_cat, dummy_Heating)
df_housing_cat<- cbind(df_housing_cat, dummy_HeatingQC)
df_housing_cat<- cbind(df_housing_cat, dummy_Electrical)
df_housing_cat<- cbind(df_housing_cat, dummy_KitchenQual)
df_housing_cat<- cbind(df_housing_cat, dummy_Functional)
df_housing_cat<- cbind(df_housing_cat, dummy_GarageType)
df_housing_cat<- cbind(df_housing_cat, dummy_GarageFinish)
df_housing_cat<- cbind(df_housing_cat, dummy_GarageQual)
df_housing_cat<- cbind(df_housing_cat, dummy_GarageCond)
df_housing_cat<- cbind(df_housing_cat, dummy_PavedDrive)
df_housing_cat<- cbind(df_housing_cat, dummy_SaleType)
df_housing_cat<- cbind(df_housing_cat, dummy_SaleCondition)
options(max.print = 100000)
```

Use Backward Elimination methods and StepAIC function to deduce the variables that are important for running the model

```
library(gbm)

## Loading required package: survival

## Loading required package: lattice

## Loading required package: splines

## Loading required package: parallel

## Loaded gbm 2.1.3

final_gbm=gbm(SalePrice ~ MSSubClass + MSZoning + LotFrontage +
                LotArea + Street + Utilities + OverallQual + OverallCond +
                YearBuilt + YearRemodAdd + MasVnrArea + BsmtFinSF1 +
BsmtFinSF2 +
                BsmtUnfSF + X1stFlrSF + X2ndFlrSF + BedroomAbvGr +
KitchenAbvGr +
                TotRmsAbvGrd + GarageCars + GarageArea + WoodDeckSF +
ScreenPorch +
                PoolArea + LandContourHLS + LandContourLow + LandContourLvl +
                LotConfigCulDSac + LotConfigFR2 + LandSlopeMod + LandSlopeSev
+
                NeighborhoodClearCr + NeighborhoodCollgCr +
NeighborhoodCrawfor +
                NeighborhoodEdwards + NeighborhoodGilbert +
NeighborhoodMitchel +
                NeighborhoodNAmes + NeighborhoodNoRidge + NeighborhoodNPkVill
+
                NeighborhoodNridgHt + NeighborhoodNWAmes +
NeighborhoodOldTown +
                NeighborhoodSawyer + NeighborhoodStoneBr + NeighborhoodTimber
+
```

```
                Condition1Norm + Condition1RRAe + Condition2PosA +
Condition2PosN +
                Condition2RRAe + BldgTypeDuplex + BldgTypeTwnhs +
BldgTypeTwnhsE +
                HouseStyle1.5Unf + HouseStyle1Story + HouseStyle2.5Fin +
                HouseStyleSFoyer + HouseStyleSLvl + RoofStyleShed +
RoofMatlCompShg +
                RoofMatlMembran + RoofMatlMetal + RoofMatlRoll +
RoofMatlTar.Grv +
                RoofMatlWdShake + RoofMatlWdShngl + Exterior1stHdBoard +
                Exterior1stPlywood + Exterior2ndImStucc + MasVnrTypeNone +
                MasVnrTypeStone + ExterQualGd + ExterQualTA + ExterCondGd +
                FoundationWood + BsmtQualFa + BsmtQualGd + BsmtQualTA +
BsmtCondTA +
                BsmtFinType1GLQ + BsmtExposureGd + BsmtExposureNo +
HeatingQCGd +
                HeatingQCTA + KitchenQualFa + KitchenQualGd + KitchenQualTA +
                FunctionalSev + FunctionalTyp + GarageFinishRFn +
GarageQualFa +
                GarageQualGd + GarageQualPo + GarageQualTA + GarageCondFa +
                GarageCondGd + GarageCondPo + GarageCondTA + SaleTypeCon +
                SaleTypeConLD + SaleTypeNew + SaleConditionNormal +
Exterior1stBrkFace +
                Exterior1stMetalSd + MasVnrTypeBrkFace,data =
df_housing_cat,distribution = "gaussian",
                n.trees = 10000,shrinkage = 0.01, interaction.depth = 4)
```

Now use this predictor for predicting test data value. Follow the same procedure on removing, Adding the impute values by creating new dataframe df_housing_test_cat. Only additonal operation to do is to make sure all variables that are used in the prediction are available. if not avaialble add the missing variable and set it to 0 ` ``` Let us train for test dataset

```
housing_test<-read.csv("test.csv")
df_housing_test=housing_test
#find missing value by columns
colSums(is.na(df_housing_test))
```

```
##           Id    MSSubClass      MSZoning   LotFrontage       LotArea
##            0             0             4           227             0
##       Street         Alley      LotShape   LandContour     Utilities
##            0          1352             0             0             2
##    LotConfig     LandSlope  Neighborhood    Condition1    Condition2
##            0             0             0             0             0
##     BldgType    HouseStyle   OverallQual   OverallCond     YearBuilt
##            0             0             0             0             0
##  YearRemodAdd     RoofStyle      RoofMatl   Exterior1st   Exterior2nd
##            0             0             0             1             1
##   MasVnrType    MasVnrArea     ExterQual     ExterCond    Foundation
##           16            15             0             0             0
```

```
##      BsmtQual      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1
##            44            45            44            42             1
##  BsmtFinType2    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating
##            42             1             1             1             0
##     HeatingQC    CentralAir    Electrical     X1stFlrSF     X2ndFlrSF
##             0             0             0             0             0
##  LowQualFinSF     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath
##             0             0             2             2             0
##      HalfBath  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd
##             0             0             0             1             0
##    Functional    Fireplaces   FireplaceQu    GarageType   GarageYrBlt
##             2             0           730            76            78
##  GarageFinish    GarageCars    GarageArea    GarageQual    GarageCond
##            78             1             1            78            78
##    PavedDrive    WoodDeckSF   OpenPorchSF EnclosedPorch    X3SsnPorch
##             0             0             0             0             0
##   ScreenPorch      PoolArea        PoolQC         Fence   MiscFeature
##             0             0          1456          1169          1408
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##             0             0             0             1             0
```

```
#Remove columns with 90% NA values
df_housing_test=subset(df_housing_test,select = -
c(PoolQC,Fence,MiscFeature,Alley,FireplaceQu))

colSums(is.na(df_housing_test))
```

```
##            Id    MSSubClass      MSZoning   LotFrontage       LotArea
##             0             0             4           227             0
##        Street      LotShape   LandContour     Utilities     LotConfig
##             0             0             0             2             0
##     LandSlope  Neighborhood    Condition1    Condition2      BldgType
##             0             0             0             0             0
##    HouseStyle   OverallQual   OverallCond     YearBuilt  YearRemodAdd
##             0             0             0             0             0
##     RoofStyle      RoofMatl   Exterior1st   Exterior2nd    MasVnrType
##             0             0             1             1            16
##    MasVnrArea     ExterQual     ExterCond    Foundation      BsmtQual
##            15             0             0             0            44
##      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1  BsmtFinType2
##            45            44            42             1            42
##    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating     HeatingQC
##             1             1             1             0             0
##    CentralAir    Electrical     X1stFlrSF     X2ndFlrSF  LowQualFinSF
##             0             0             0             0             0
##     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath      HalfBath
##             0             2             2             0             0
##  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd    Functional
##             0             0             1             0             2
##    Fireplaces    GarageType   GarageYrBlt  GarageFinish    GarageCars
```

```
##            0           76           78           78            1 
##   GarageArea   GarageQual   GarageCond   PavedDrive   WoodDeckSF 
##            1           78           78            0            0 
##  OpenPorchSF EnclosedPorch   X3SsnPorch  ScreenPorch     PoolArea 
##            0            0            0            0            0 
##      MiscVal       MoSold       YrSold     SaleType SaleCondition 
##            0            0            0            1            0
```

```
df_housing_test$LotFrontage
```

```
##    [1]  80  81  74  78  43  75  NA  63  85  70  26  21  21  24  24 102  94
##   [18]  90  79 110 105  41 100  43  67  63  60  73  92  84  70  70  39  85
##   [35]  88  25  39  30  24  24  NA  NA  57  68  80  NA  80  NA  80  80  90
##   [52]  88  NA  98  68 120  75  70  70  NA  87  80  60  60 119  70  65  60
##   [69]  81  80  60  56  69  50  69  NA  68  60  50 100  60  53  NA  50  50
##   [86]  50  53  50  52  52  51  57  60  52 100  72  60  65  NA  60  72  65
##  [103]  65  NA  86  NA  94  NA 124  65  50  75  44  NA  83  87  64  82  82
##  [120]  NA  38  68  80  75  NA  67  68  60  89  65  64  67  NA  60  51  78
##  [137]  78  85  35  35  58  50  66  44  85  74  NA  88  73  73  85  93  NA
##  [154]  31  21  21  21  50  76  70  63  68  76  74  74  85  88  NA  60  28
##  [171]  61  57  57  60  NA  58  85  NA  80  NA  80  70  NA  NA  NA  78  85
##  [188]  NA  NA  60  60  21  21  24  24  24  24  24  24  NA 110  95  95 105
##  [205]  95 129  59  87  77 102  90 110  96  70  47  34  80 100 117  44  48
##  [222] 129  48  63  57  43  59  62  61  NA  NA  NA  61  42  62  NA  64 106
##  [239]  NA  79  NA  86  78  85  76  85  90  72 112  75  84  65  85  68  65
##  [256]  80  63  63  96  76  63  63  60  61  43  70  50  70  NA  75  63  NA
##  [273]  NA  NA  NA  65  NA  NA  32  NA  NA  34  35  NA 110  80  NA  80  80
##  [290]  75  NA  62  80  80  NA  60  65  NA  NA  85 115  NA  85  68  90  92
##  [307]  80  73  NA  66  70  70  80  76  53  67  80  60  75  78  60  53  60
##  [324]  80  60  60  60  90  60  60  81  83  77  62  90  80  60  71  60  80
##  [341]  60  60  76  75  80  68  57  90  90  57  63  56  50  62  50  60  60
##  [358]  70  60  NA  60  60  72  NA  50  60  51  51  50  57  68  50  57  NA
##  [375]  41  60  86  60  50  60  NA  60  75  NA  88  88  NA  NA  NA  68  50
##  [392]  62  42  74  66  85 120  64  64  64  64  NA  NA  NA  84  65  71  77
##  [409]  64  95  78  79  NA  65  65  70  65  75  NA  76  90  NA  70  NA  90
##  [426]  NA  70  85  85  80  35  64  70  65  70  45  70  43  64  53  60  80
##  [443]  NA  70  90  78 100  24  24  24  NA  NA  50  60  44 109  75  75  72
##  [460]  82 113  79  NA 125  75  85  75  83  50  62  70  62  70  72  65  59
##  [477]  NA  53  45  39  73  NA  65 101  53  NA  60  NA  63  NA  56  85  90
##  [494]  80  75  NA  80  NA  60  68  63  21  21  21  24  24  98 105 104 108
##  [511]  96 102  74  85 106  92 130 112  58 135  89  48  48  36  NA  NA  53
##  [528]  80  NA  NA  55  71  NA  41  77  84  NA 136  97  NA  91  74  73  80
##  [545]  87  72  85  62  68  67  63  81  65  50  43  65  75  70  75  65  60
##  [562]  95  70  NA 105  37  30  30  24  NA  NA  42  35  24  79  24  24  36
##  [579]  22  NA 103  NA  NA  85  NA  75  73  65  72  NA  74  90  NA  50  80
##  [596]  80  63  90  74  82  90  75  60 102  NA  95  71  76  60  45  60  60
##  [613]  78  76  NA  60  80  80  60  64  60 113  60  60  69  56  57  80  60
##  [630]  60  63  63  81  60  60  60  44  75  62 103  69  53  69  60  60  60
##  [647]  60  65  52  55  NA  NA  50  59  50  50  50  50  50  60  99  52  NA
##  [664]  52  51  60  57  63  NA  60  60  NA  60  62  60  60  82  NA  80  68
```

```
##  [681]  NA  NA  NA  60  94  88  NA  63  NA  82  50  85  68  50  50  NA  80
##  [698]  NA  85  65  65  91  91  NA  NA  65  NA  NA  68  75  NA  NA  40  NA
##  [715]  NA  NA  41  96  NA  83  75  85  74  75  64  65  NA  72 123  65  74
##  [732]  56  60  57  68  62  67  60  64  66  63  54  NA  NA  NA  79 100  70
##  [749]  56  24  50  60  58  75  60  75  80  60  52  60  44  44  NA  NA  76
##  [766]  74  42  74 107  73  81  75  93  82  NA  79  85  97  77  32 150  NA
##  [783]  41  21  21  NA  21  59  60  74  NA  85  56  NA  NA  63  60  NA  NA
##  [800] 160  38  35  98  52 195  61  85  81  NA  78  93  61  79  80 128  64
##  [817]  80  63  66  NA  33  26  21  21  24  65  96  91 110 107 110 105 107
##  [834] 118  59 134  82  94  99 110  NA  70  71  92  34  34  41  34  48  48
##  [851]  48  59  65  58  62  63  88  72  64  64  NA  53  65  87  59  NA  63
##  [868]  NA  58  59 100  89  74  83  88  82  75  91  76  98  85  74  70  70
##  [885]  75  72  75  72  81 112  84 100  85  63  63  60  65  NA  50  64  84
##  [902]  NA  44  43  45  80  32  40  30  24  24  73  NA 106  50  80  94  78
##  [919]  80  NA 130 108  80  78  88  80  70  NA  74  NA  76  70  80  80  80
##  [936]  NA  NA  85  50  51  50  80  63  NA 100  73  65  60  70  80  72  70
##  [953]  75  60  70  60  68  71  55  70  75  NA  60  NA 113  60  60  93  75
##  [970]  66  60  NA  70  70  80  70  60  66  60  50  56  56  60 120  50  57
##  [987]  NA  53  35  56  60  75  52  76  55  55  50  50  51  NA  52  52  51
## [1004]  47  60  60  NA  40  40 120  60  60  52  60 107  59  75  75  62  65
## [1021]  NA  70  86  NA  NA  81  99  70  91  85  NA  NA  NA  NA  84 102  70
## [1038]  60  NA  39  58  60  50 104  75  71  93  66  75  73  64  78  NA 155
## [1055]  NA  57  60  70  47  43  68  NA  97  59  72  45  39  75  60  60  NA
## [1072]  70  65  73  NA  NA  73  70  65  64  64  64  75  67  74  80  65  NA
## [1089]  60 128  35  64  74  52  62  60  60  60  54  51  63  53  53  60 126
## [1106] 110  79  NA  NA  24  NA  35  NA  70  50  50  50  46  64  75  65  60
## [1123]  59  NA  80  44  NA  NA  82  46 149  67  68  42  NA  80  NA  NA  85
## [1140] 200  62  21  NA  21  21  72  NA  61  68  50 124  NA  65  62  NA  85
## [1157]  NA  NA  90  60  NA  54  50  42  68  NA  30  59  60  63  82  92  60
## [1174]  90  NA  81  NA  75  81  80  60  26  24  24  21  21  21  53  65  24
## [1191]  24  72 110 108 120 120  82 103  82  82 121 131  NA  48  61  48  65
## [1208]  65 102  96  75  43  NA  NA  43  59  NA  NA  84  83  NA  83 114  NA
## [1225]  NA  75  49  85  72 100  65  74  91  63  65  74  70  70  50  50  NA
## [1242]  70  68  NA  NA  65  NA  NA  NA  NA 100 105  34  24  NA  NA 114  60
## [1259]  79  78  80  72  78  70  NA  80  80  NA  85  80  60  68  80  89  80
## [1276]  79  82  NA  NA 109  70  NA 125  72  70  66  75  55  65  80  85 118
## [1293]  70  94  50  60  60  60  60  60  83  77  80  86  NA  60  60  75  70
## [1310]  74  70  70  62  60  60  60  58  66  56  56  50  60  42  50  50  52
## [1327]  56  48  70  33  65  63  69  51  51  50  90  60  60  50  60  82 120
## [1344] 100  55  50  50  NA  80  75  NA  NA  NA  75  NA  NA  87  72  NA  75
## [1361]  62 114  60  78  80  70  75  88  73 133  64  90  78  91  78  78  80
## [1378]  95  65  NA  68  72  50  42  60  NA  45  NA  70  67  NA  90  90  37
## [1395]  70  74  70  65  67  38  73  62  75  90  35  24  56  72  62  60  45
## [1412]  60  76  60  58  43  69  54  84  51  66  80  81  70  70  78  50  50
## [1429]  61  50  75  69  50  60  41  44  69  65  70 140  NA  NA  95  88 125
## [1446]  78  41  58  NA  21  21  80  21  21  21  21 160  62  74
```

```
mean_Lfrontage=mean(df_housing_test$LotFrontage,na.rm = TRUE)
mean_Lfrontage
```

```
## [1] 68.58036
```

```
df_housing_test$LotFrontage[is.na(df_housing_test$LotFrontage)]=mean_Lfrontag
e
```

```
df_housing_test$GarageYrBlt
```

```
##    [1] 1961 1958 1997 1998 1992 1993 1992 1998 1990 1970 1999 1971 1997
##   [14] 1975 1975 2009 2009 2005 2005 2003 2002 2006 2005 2006 2004 2004
##   [27] 1998 2005 2009 2005 2004 1920 1974 1993 1992 2004 2004 2004 2004
##   [40] 2005 2000 2003 2010 2000 2002 1967 1993 1978 1971 1966 1966 1967
##   [53] 1964   NA 1994 1949 1966 1958 2003 1959 1959 1956 1956 1952 1955
##   [66] 1958 1989 1950 1960 1963 1900   NA 1957 1938 1948 1962 1928 1930
##   [79] 2003   NA 1970 1950 1928 1926 1939 1973 1942 1948 1979 1930 1923
##   [92] 1915   NA 1920 1959 1917   NA 1940   NA 1910   NA 1966 1969 1978
##  [105] 1968 1977 1945 1978 1938 1987 1947 1954 2009 1964 1987 2000 2009
##  [118] 1957 1998 1997 1977 1977 2003 1997 2003 1945 1954 1968 1956 1975
##  [131]   NA 1979 1939   NA   NA 1941 1950 1994 1989 1989 1951 1950 1896
##  [144] 2004 1998 1977 1976 2008 2010 2007 1965 2004 2001 1973   NA   NA
##  [157] 1972 1971 1984 1985 1993 1969 1994 1993 1956 1974 1997 2003 1996
##  [170] 2004 1998 1995 1998 1998 1994 1993 1977 1978 1978 1980 1978 2003
##  [183] 2000 2002 1975 1974 1975 1970 1971 2001 1986 1973 1972 1976 1975
##  [196] 1977 1978 1978 1976 1966 2007 2009 2008 2007 2008 2004 2007 2008
##  [209] 2006 2008 2003 2003 2003 2006 2005 2005 2008 2004 2003 2008 2008
##  [222] 2002 2003 2005 2005 2005 2004 2004 2004 2003 2003 2002 2004 2000
##  [235] 1999 1999 1999 2000 1994 1995 1993 2008 2008 2007 2006 2005 2008
##  [248] 2008 2008 2006 2006 2009 2006 2003 2003 2007 2006   NA 2004 2004
##  [261] 2005   NA 2004 2008 1997 1992 1990 1994 1986 1981 1969 1982 1961
##  [274] 1965 1963 1962 1980 1991 2004 2008 2008 2000 1999 1977 1981 1976
##  [287] 1974 1967 1969 1969 1977 1967 1967 1974 1971 1988 1960 1982 1956
##  [300] 1961 1964 1965 1961 1955 1967 1961 1966 1956 1960 1959 1956 1955
##  [313] 1956 1958 1954 1951 1945 1952 1953 1948 1950 1958 1939 1940 1987
##  [326] 1954 2008   NA 1980 1959 1969 1963 1967 1985 1957 1958 1989 1958
##  [339] 1952 1959 1949 1994 1964 1978 1963 1920 1920 1959   NA 1939   NA
##  [352]   NA 1950 1920 1965 1963 1974 1930 1917   NA 1920 1950   NA 1923
##  [365] 1955 1924 1926 1938 1982 1930 1915   NA 1927 1915   NA 1927   NA
##  [378] 1915 1946   NA 1960 1934 1984 1978 1961 1960 1956   NA 1980 1956
##  [391] 1946 1954 1984 1990 1983 1993 1900 1979 1979 1979 1979 2000 2000
##  [404] 2000 2009 2008 2008 2007 2007 2008 2005 2005 1992 1995 1998 1998
##  [417] 2002 2001 1978 1979 2002 2003 2002 2001 1999 2002 1997 2007 2007
##  [430] 1968 2005 1959 1950   NA 1956 1940 1938 1926 1916 1918 1961 1960
##  [443] 1940 1954 1960 1949 1954 1980 1980 1980 1986 1971 1998 1940 2007
##  [456] 1975 2000 1977 1977 1991 2008 2008 1980 1987 2003 2007 1968 1969
##  [469] 1998 1993 1998 2001 1969 1997 1995 1998 1996 1996 1997 1992 1998
##  [482] 1991 1989 2005 2004 1952 2007 1950 1988 1983 1978 1979 1976 1980
##  [495] 1969 1978 1976 1996 1982 1969 1977 1973 1975 1972 1975 1977 2007
##  [508] 2007 2007 2006 2005 2005 2007 2004 2003 2003 2001 2003 2008 2005
##  [521] 2008 2007 2007 2003 2003 2004 2007 2004 2002 2003 2008 2000 2002
##  [534] 1999 1999 1997 2000 1998 1996 1995 1993 2006 2007 2007 2006 2008
##  [547] 2007 2008 2003 2003   NA 2007 1995 1993 1994 2001 1992 1963 1962
```

```
##  [560] 1970 1963 1974 1972 1990 1993 2004 2005 2007 1999 2000 2001 2001
##  [573] 1999 1999 2001 1999 2000 1998 1995 1977 1976 2002 1969 1968 1967
##  [586] 1965 1968 1978 1987 1971 1956 1961 1960 1937 1960 1950 1953 1966
##  [599] 1957 1959 1958 1956 1952 1971 1953 1957 1957 1958 1948 1932 1997
##  [612] 1968 1990 1958 1960 1972 1959 1962 1994 1954 1954   NA 1955 1954
##  [625] 1963 2008 1948 1910 1950 1915   NA 1958 1920   NA 1940 1930   NA
##  [638] 1959 1949   NA 1950 1935 1961 1930   NA 1920 1950 1959 1959 1992
##  [651] 1945 1950 1941 1926 1940 1924 2004 1939 1926 1920 1946 1990 1925
##  [664] 1939 1960 1970   NA 1910 1930 1952 1938 1950 1993 1985 1997   NA
##  [677] 1947 1996 1978 1967 1978 1967 1984 1920 1963 1956 1960 1973 1979
##  [690] 1979 1948   NA 1956   NA 1962 1995 1994 1993 1996 2007 2007 2008
##  [703] 2008 1995 1966 1997 1997 1997 2000 2000 1978 1985 1975 2001 2001
##  [716] 2002 2003 1999 2004 1998 2004 2007 2007 1966 1976 1991 1977 1976
##  [729] 1959   NA   NA   NA   NA   NA 1953 1954 1923 1921 1930 2001 1994
##  [742] 1925 1980 1937 1938 1951 1935 1994 1956 1980 1926 1940   NA 1967
##  [755] 1934 1958 1952 1895 1910 1920 2007 2004 1996 1996 1976 1991 1986
##  [768] 2007 2007 2007 2008 1989 1986 2003 1999 2007 2005 1997   NA 1997
##  [781] 1964 1975 1976 1973 1973 1968   NA 1983 1982 1984 1900 1971 1997
##  [794] 1994 2000 1996 1999 1992 1993 1964 1988 1990 2005 2005 1969 2006
##  [807] 2006 2007 1984 1981 1978 1979 1984 1979 1971 1976 1974 1988 1970
##  [820] 1961 2001 1997 1973 1973 1978 1974 2007 2006 2007 2007 2006 2005
##  [833] 2007 2005 2007 2007 2007 2006 2004 2003 2007 2005 2005 2006 2005
##  [846] 2005 2005 2006 2005 2006 2003 2007 2005 2007 2006 2007 2005 2007
##  [859] 2007 2007 2003 2007 2004 2004 2006 2003 2003 2002 2000 1999 1998
##  [872] 1998 1998 1998 1992 1996 2007 2007 2006 2007 2007 2005 2005 2006
##  [885] 2006 2007 2007 2007 2007 2007 2007 2006 2004   NA   NA 2004 2006
##  [898] 1993 1980 1979 1991 1990 1974 1973 2004 2006 2006 2006 2007 1999
##  [911] 2000 1999 2003 1998 1994 1980 1981 1968 1970 1969 1968 1972 1993
##  [924] 1993 1966 1963 1967 1964 1966 1961 1985 1965 1966 1965 1964 1964
##  [937] 1959 1975   NA   NA 1963 1955 1968 1966 1961 1957 1964 1994 1960
##  [950] 1957 1957 1955 1955 1962 1958 1952 1953 1956 2002 1955 1953 1952
##  [963]   NA 1953 2007 1978   NA 1963 1961 1968 1950 1959 1958 1960 1965
##  [976] 1961 1962 1962 1981 1980 1922 1920 1940 2000 1930 1935 1992 1927
##  [989] 1920 1979 2004 1950 1956 1957 1969 1950 1939 1939 1939 1968 1939
## [1002] 1930 1926 1950 1977 1965 1979 1920 1920 1963 1950 2006 1958 1910
## [1015] 1937 1942 1963 1964 1964 1970 1966 1989 1968 1972 1966 1956 1946
## [1028] 1940 1954 1958 1952 1984 1996 1953 1946 1954 1954 1958 1958 1984
## [1041] 1951 1951 1920 1984 1994 2007 2006 2007 2005 2005 2005 2005 1988
## [1054] 1976 1995 1997 1995 1996 1999 1998 2001 2000 1974 1979 1977 1977
## [1067] 1977 1975 1977 1975 2000 2003 2002 1994 2001 1996 1999 2007 2005
## [1080] 2006 2006 2005 1967 2002 1975 1960 1976 1979 2005 2008 2005 1959
## [1093]   NA   NA 1920 1959 1996   NA 1973 1994 1930 1992 1926 1927 1951
## [1106] 1930 1966 1960 1968 1980 1996 1988 1971 1986 1965   NA   NA 1922
## [1119] 1950   NA 1985 1930 2006 1979 2002 2002 1991 1975 1974 1987 1958
## [1132] 2007 2207 1985 2001 2002 1996 2003 2006 1953 1996 1972 1970   NA
## [1145] 1976 1977 1977 1977 1977   NA 1961 1976 1983 1984 1954 1956 1957
## [1158] 1957 1969 1997 1995 1996 2005 2006 1994 1993 1987 2005 2006 2006
## [1171] 2006 2005 2005 2005 1980 1978 1976 2005 1975 1974 1995 2002 1973
## [1184] 1973 1973 1972 1972 1975 1974 1976 1976 2006 2004 2006 2005 2005
## [1197] 2005 2005 2006 2005 2005 2004 2004 2005 2006 2006 2005 2006 2006
```

```
## [1210] 2004 2006 2006 2002 2004 2005 2004 2000 2006 1998 2000 2000 1995
## [1223] 1993 1994 1993 2005 2006 2006 2006 2006 2005   NA 2005   NA 2005
## [1236] 2005 1997 1992 1990 1991 1994 1977 1977 1972 1965 1968 1990 1965
## [1249]   NA 1967 1974 1992 2004 2005 2000 2003 1997 2001 1972 1967 1968
## [1262] 1968 1968 1968 1966 1967 1965 1964 1964 1960 1949 1947 1961 1952
## [1275] 1951 1949 1954 1967 1964 1963 1957 1958 1956 1956 1954 1952 1951
## [1288] 1993 1956 1955 1951 1941 2001 1977 1953 1936 1967 1900 1974 1935
## [1301] 1963 1962 1961 1959 1962 1950 1950   NA 1961 1962 1962   NA 1900
## [1314] 1948 1920 1956 1985 1930 1993 1995 1925 1952 1930 1976 1976 1921
## [1327] 1945 1938 1910   NA 2002   NA 1943 1930 1930 1925 1962 1924 2001
## [1340]   NA 1930 1951 1964 1950 1939 1936 2004 1967 1964 1966 1967 1979
## [1353] 1977 1977 1924 1967 1993 1973 1980 1954 1942 1928 1993 1962 2002
## [1366] 1955 1953 1989 1993 2005 2005 2006 2005 2005 2005 2005 1968 1996
## [1379] 1995 1998 1999 1999 1977 1989 1977 2002 1999 2002 2002 1999 1997
## [1392] 1998 1995 2003 2003 2006 2005 2005 1950   NA 1969 2003   NA 2003
## [1405] 2005 2004 1948 1974 1924 1956   NA 1922 1910 1938 1945 1926 1920
## [1418] 1919 1939 1941 1937 1940 1963 1963 1930 1950 1942 1950   NA 1925
## [1431] 1957   NA   NA   NA 2005 2004 1979 1978 2001 1975 1958 2000 2005
## [1444] 2005 1951 1997 1977 1968 1970   NA 1972 1969 1970   NA   NA 1970
## [1457] 1960   NA 1993
```

```
mean_GarageYrBlt=mean(df_housing_test$GarageYrBlt,na.rm = TRUE)
mean_GarageYrBlt
```

```
## [1] 1977.721
```

```
df_housing_test$GarageYrBlt[is.na(df_housing_test$GarageYrBlt)]=mean_GarageYr
Blt
colSums(is.na(df_housing_test))
```

```
##            Id    MSSubClass      MSZoning   LotFrontage       LotArea
##             0             0             4             0             0
##        Street      LotShape   LandContour     Utilities     LotConfig
##             0             0             0             2             0
##     LandSlope  Neighborhood    Condition1    Condition2      BldgType
##             0             0             0             0             0
##    HouseStyle   OverallQual   OverallCond     YearBuilt  YearRemodAdd
##             0             0             0             0             0
##     RoofStyle      RoofMatl   Exterior1st   Exterior2nd    MasVnrType
##             0             0             1             1            16
##    MasVnrArea     ExterQual     ExterCond    Foundation      BsmtQual
##            15             0             0             0            44
##      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1  BsmtFinType2
##            45            44            42             1            42
##    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating     HeatingQC
##             1             1             1             0             0
##    CentralAir    Electrical     X1stFlrSF     X2ndFlrSF  LowQualFinSF
##             0             0             0             0             0
##     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath      HalfBath
##             0             2             2             0             0
##  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd    Functional
```

```
##             0             0             1             0             2
##    Fireplaces    GarageType   GarageYrBlt  GarageFinish    GarageCars
##             0            76             0            78             1
##    GarageArea    GarageQual    GarageCond    PavedDrive    WoodDeckSF
##             1            78            78             0             0
##   OpenPorchSF EnclosedPorch    X3SsnPorch   ScreenPorch      PoolArea
##             0             0             0             0             0
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##             0             0             0             1             0
```

```
df_housing.imp<-missForest(df_housing_test)
```

```
##   missForest iteration 1 in progress...
```

```
## Warning in randomForest.default(x = obsX, y = obsY, ntree = ntree, mtry =
## mtry, : The response has five or fewer unique values. Are you sure you
want
## to do regression?
```

```
## Warning in randomForest.default(x = obsX, y = obsY, ntree = ntree, mtry =
## mtry, : The response has five or fewer unique values. Are you sure you
want
## to do regression?
```

```
## done!
##   missForest iteration 2 in progress...
```

```
## Warning in randomForest.default(x = obsX, y = obsY, ntree = ntree, mtry =
## mtry, : The response has five or fewer unique values. Are you sure you
want
## to do regression?
```

```
## Warning in randomForest.default(x = obsX, y = obsY, ntree = ntree, mtry =
## mtry, : The response has five or fewer unique values. Are you sure you
want
## to do regression?
```

```
## done!
##   missForest iteration 3 in progress...
```

```
## Warning in randomForest.default(x = obsX, y = obsY, ntree = ntree, mtry =
## mtry, : The response has five or fewer unique values. Are you sure you
want
## to do regression?
```

```
## Warning in randomForest.default(x = obsX, y = obsY, ntree = ntree, mtry =
## mtry, : The response has five or fewer unique values. Are you sure you
want
## to do regression?
```

```
## done!
##   missForest iteration 4 in progress...

## Warning in randomForest.default(x = obsX, y = obsY, ntree = ntree, mtry =
## mtry, : The response has five or fewer unique values. Are you sure you
want
## to do regression?

## Warning in randomForest.default(x = obsX, y = obsY, ntree = ntree, mtry =
## mtry, : The response has five or fewer unique values. Are you sure you
want
## to do regression?

## done!
##   missForest iteration 5 in progress...

## Warning in randomForest.default(x = obsX, y = obsY, ntree = ntree, mtry =
## mtry, : The response has five or fewer unique values. Are you sure you
want
## to do regression?

## Warning in randomForest.default(x = obsX, y = obsY, ntree = ntree, mtry =
## mtry, : The response has five or fewer unique values. Are you sure you
want
## to do regression?

## done!

df_housing_test=df_housing.imp$ximp
df_housing_test<-na.omit(df_housing_test)
colSums(is.na(df_housing_test))

##            Id    MSSubClass      MSZoning   LotFrontage       LotArea
##             0             0             0             0             0
##        Street      LotShape   LandContour     Utilities     LotConfig
##             0             0             0             0             0
##     LandSlope  Neighborhood    Condition1    Condition2      BldgType
##             0             0             0             0             0
##    HouseStyle   OverallQual   OverallCond     YearBuilt  YearRemodAdd
##             0             0             0             0             0
##     RoofStyle      RoofMatl   Exterior1st   Exterior2nd    MasVnrType
##             0             0             0             0             0
##    MasVnrArea     ExterQual     ExterCond    Foundation      BsmtQual
##             0             0             0             0             0
##      BsmtCond  BsmtExposure  BsmtFinType1    BsmtFinSF1  BsmtFinType2
##             0             0             0             0             0
##    BsmtFinSF2     BsmtUnfSF   TotalBsmtSF       Heating     HeatingQC
##             0             0             0             0             0
##    CentralAir    Electrical     X1stFlrSF     X2ndFlrSF  LowQualFinSF
##             0             0             0             0             0
##     GrLivArea  BsmtFullBath  BsmtHalfBath      FullBath      HalfBath
```

```
##             0             0             0             0             0
##  BedroomAbvGr  KitchenAbvGr   KitchenQual  TotRmsAbvGrd    Functional
##             0             0             0             0             0
##    Fireplaces    GarageType   GarageYrBlt  GarageFinish    GarageCars
##             0             0             0             0             0
##    GarageArea    GarageQual    GarageCond    PavedDrive    WoodDeckSF
##             0             0             0             0             0
##   OpenPorchSF EnclosedPorch    X3SsnPorch   ScreenPorch      PoolArea
##             0             0             0             0             0
##       MiscVal        MoSold        YrSold      SaleType SaleCondition
##             0             0             0             0             0
```

```
#Category variable for factor level 2
levels(df_housing_test$Street)<-c(1,0)
df_housing_test$Street<-
as.numeric(levels(df_housing_test$Street))[df_housing_test$Street]

levels(df_housing_test$Utilities)<-c(1,0)
df_housing_test$Utilities<-
as.numeric(levels(df_housing_test$Utilities))[df_housing_test$Utilities]

levels(df_housing_test$CentralAir)<-c(1,0)
df_housing_test$CentralAir<-
as.numeric(levels(df_housing_test$CentralAir))[df_housing_test$CentralAir]

#Create dummy variable for remaining factor variable
dummy_MSZoning  <- data.frame(model.matrix( ~MSZoning , data =
df_housing_test))
dummy_MSZoning  <- dummy_MSZoning[,-1]
dummy_LotShape<- data.frame(model.matrix( ~LotShape, data = df_housing_test))
dummy_LotShape<-dummy_LotShape[,-1]
dummy_LandContour<- data.frame(model.matrix( ~LandContour, data =
df_housing_test))
dummy_LandContour<-dummy_LandContour[,-1]
dummy_LotConfig<- data.frame(model.matrix( ~LotConfig, data =
df_housing_test))
dummy_LotConfig<-dummy_LotConfig[,-1]
dummy_LandSlope<- data.frame(model.matrix( ~LandSlope, data =
df_housing_test))
dummy_LandSlope<-dummy_LandSlope[,-1]
dummy_Neighborhood<- data.frame(model.matrix( ~Neighborhood, data =
df_housing_test))
dummy_Neighborhood<-dummy_Neighborhood[,-1]
dummy_Condition1<- data.frame(model.matrix( ~Condition1, data =
df_housing_test))
dummy_Condition1<-dummy_Condition1[,-1]
dummy_Condition2<- data.frame(model.matrix( ~Condition2, data =
df_housing_test))
dummy_Condition2<-dummy_Condition2[,-1]
dummy_BldgType<- data.frame(model.matrix( ~BldgType, data = df_housing_test))
```

```
dummy_BldgType<-dummy_BldgType[,-1]
dummy_HouseStyle<- data.frame(model.matrix( ~HouseStyle, data =
df_housing_test))
dummy_HouseStyle<-dummy_HouseStyle[,-1]
dummy_RoofStyle<- data.frame(model.matrix( ~RoofStyle, data =
df_housing_test))
dummy_RoofStyle<-dummy_RoofStyle[,-1]
dummy_RoofMatl<- data.frame(model.matrix( ~RoofMatl, data = df_housing_test))
dummy_RoofMatl<-dummy_RoofMatl[,-1]
dummy_Exterior1st<- data.frame(model.matrix( ~Exterior1st, data =
df_housing_test))
dummy_Exterior1st<-dummy_Exterior1st[,-1]
dummy_Exterior2nd<- data.frame(model.matrix( ~Exterior2nd, data =
df_housing_test))
dummy_Exterior2nd<-dummy_Exterior2nd[,-1]
dummy_MasVnrType<- data.frame(model.matrix( ~MasVnrType, data =
df_housing_test))
dummy_MasVnrType<-dummy_MasVnrType[,-1]
dummy_ExterQual<- data.frame(model.matrix( ~ExterQual, data =
df_housing_test))
dummy_ExterQual<-dummy_ExterQual[,-1]
dummy_ExterCond<- data.frame(model.matrix( ~ExterCond, data =
df_housing_test))
dummy_ExterCond<-dummy_ExterCond[,-1]
dummy_Foundation<- data.frame(model.matrix( ~Foundation, data =
df_housing_test))
dummy_Foundation<-dummy_Foundation[,-1]
dummy_BsmtQual<- data.frame(model.matrix( ~BsmtQual, data = df_housing_test))
dummy_BsmtQual<-dummy_BsmtQual[,-1]
dummy_BsmtCond<- data.frame(model.matrix( ~BsmtCond, data = df_housing_test))
dummy_BsmtCond<-dummy_BsmtCond[,-1]
dummy_BsmtFinType1<- data.frame(model.matrix( ~BsmtFinType1, data =
df_housing_test))
dummy_BsmtFinType1<-dummy_BsmtFinType1[,-1]
dummy_BsmtExposure<- data.frame(model.matrix( ~BsmtExposure, data =
df_housing_test))
dummy_BsmtExposure<-dummy_BsmtExposure[,-1]
dummy_BsmtFinType2<- data.frame(model.matrix( ~BsmtFinType2, data =
df_housing_test))
dummy_BsmtFinType2<-dummy_BsmtFinType2[,-1]
dummy_Heating<- data.frame(model.matrix( ~Heating, data = df_housing_test))
dummy_Heating<-dummy_Heating[,-1]
dummy_HeatingQC<- data.frame(model.matrix( ~HeatingQC, data =
df_housing_test))
dummy_HeatingQC<-dummy_HeatingQC[,-1]
dummy_Electrical<- data.frame(model.matrix( ~Electrical, data =
df_housing_test))
dummy_Electrical<-dummy_Electrical[,-1]
dummy_KitchenQual<- data.frame(model.matrix( ~KitchenQual, data =
df_housing_test))
```

```
dummy_KitchenQual<-dummy_KitchenQual[,-1]
dummy_Functional<- data.frame(model.matrix( ~Functional, data =
df_housing_test))
dummy_Functional<-dummy_Functional[,-1]
dummy_GarageType<- data.frame(model.matrix( ~GarageType, data =
df_housing_test))
dummy_GarageType<-dummy_GarageType[,-1]
dummy_GarageFinish<- data.frame(model.matrix( ~GarageFinish, data =
df_housing_test))
dummy_GarageFinish<-dummy_GarageFinish[,-1]
dummy_GarageQual<- data.frame(model.matrix( ~GarageQual, data =
df_housing_test))
dummy_GarageQual<-dummy_GarageQual[,-1]
dummy_GarageCond<- data.frame(model.matrix( ~GarageCond, data =
df_housing_test))
dummy_GarageCond<-dummy_GarageCond[,-1]
dummy_PavedDrive<- data.frame(model.matrix( ~PavedDrive, data =
df_housing_test))
dummy_PavedDrive<-dummy_PavedDrive[,-1]
dummy_SaleType<- data.frame(model.matrix( ~SaleType, data = df_housing_test))
dummy_SaleType<-dummy_SaleType[,-1]
dummy_SaleCondition<- data.frame(model.matrix( ~SaleCondition, data =
df_housing_test))
dummy_SaleCondition<-dummy_SaleCondition[,-1]

#Remove the variables from the original dataset for which dummy variables are
created
df_housing_test=subset(df_housing_test,select=-
c(LotShape,LandContour,LotConfig,LandSlope,Neighborhood,Condition1,Condition2
,BldgType,HouseStyle,RoofStyle,RoofMatl,Exterior1st,Exterior2nd,MasVnrType,Ex
terQual,ExterCond,Foundation,BsmtQual,BsmtCond,BsmtFinType1,BsmtExposure,Bsmt
FinType2,Heating,HeatingQC,Electrical,KitchenQual,Functional,GarageType,Garag
eFinish,GarageQual,GarageCond,PavedDrive,SaleType,SaleCondition))

# Combine the dummy variables to the actual dataset
df_housing_test_cat<- cbind(df_housing_test, dummy_MSZoning)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_LotShape)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_LandContour)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_LotConfig)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_LandSlope)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_Neighborhood)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_Condition1)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_Condition2)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_BldgType)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_HouseStyle)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_RoofStyle)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_RoofMatl)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_Exterior1st)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_Exterior2nd)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_MasVnrType)
```

```
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_ExterQual)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_ExterCond)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_Foundation)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_BsmtQual)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_BsmtCond)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_BsmtFinType1)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_BsmtExposure)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_BsmtFinType2)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_Heating)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_HeatingQC)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_Electrical)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_KitchenQual)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_Functional)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_GarageType)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_GarageFinish)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_GarageQual)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_GarageCond)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_PavedDrive)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_SaleType)
df_housing_test_cat<- cbind(df_housing_test_cat, dummy_SaleCondition)
options(max.print = 100000)

df_colnames=colnames(df_housing_test_cat)
df_org_stat<-
c("MSSubClass","MSZoning","LotFrontage","LotArea","Street","Utilities","Overa
llQual","OverallCond","YearBuilt","YearRemodAdd","MasVnrArea","BsmtFinSF1","B
smtFinSF2","BsmtUnfSF","X1stFlrSF","X2ndFlrSF","BedroomAbvGr","KitchenAbvGr",
"TotRmsAbvGrd","GarageCars","GarageArea","WoodDeckSF","ScreenPorch","PoolArea
","LandContourHLS","LandContourLow","LandContourLvl","LotConfigCulDSac","LotC
onfigFR2","LandSlopeMod","LandSlopeSev","NeighborhoodClearCr","NeighborhoodCo
llgCr","NeighborhoodCrawfor","NeighborhoodEdwards","NeighborhoodGilbert","Nei
ghborhoodMitchel","NeighborhoodNAmes","NeighborhoodNoRidge","NeighborhoodNPkV
ill","NeighborhoodNridgHt","NeighborhoodNWAmes","NeighborhoodOldTown","Neighb
orhoodSawyer","NeighborhoodStoneBr","NeighborhoodTimber","Condition1Norm","Co
ndition1RRAe","Condition2PosA","Condition2PosN","Condition2RRAe","BldgTypeDup
lex","BldgTypeTwnhs","BldgTypeTwnhsE","HouseStyle1.5Unf","HouseStyle1Story","
HouseStyle2.5Fin","HouseStyleSFoyer","HouseStyleSLvl","RoofStyleShed","RoofMa
tlCompShg","RoofMatlMembran","RoofMatlMetal","RoofMatlRoll","RoofMatlTar.Grv"
,"RoofMatlWdShake","RoofMatlWdShngl","Exterior1stHdBoard","Exterior1stPlywood
","Exterior2ndImStucc","MasVnrTypeNone","MasVnrTypeStone","ExterQualGd","Exte
rQualTA","ExterCondGd","FoundationWood","BsmtQualFa","BsmtQualGd","BsmtQualTA
","BsmtCondTA","BsmtFinType1GLQ","BsmtExposureGd","BsmtExposureNo","HeatingQC
Gd","HeatingQCTA","KitchenQualFa","KitchenQualGd","KitchenQualTA","Functional
Sev","FunctionalTyp","GarageFinishRFn","GarageQualFa","GarageQualGd","GarageQ
ualPo","GarageQualTA","GarageCondFa","GarageCondGd","GarageCondPo","GarageCon
dTA","SaleTypeCon","SaleTypeConLD","SaleTypeNew","SaleConditionNormal","Exter
ior1stBrkFace","Exterior1stMetalSd","MasVnrTypeBrkFace")
df_diff<-setdiff(df_org_stat,df_colnames)
for (missedvariable in df_diff) {
```

```
  df_housing_test_cat[missedvariable]=0
}
```

Now use the generated frame df_housing_test_cat for prediction and write the final output to csv file

```
Predict_price <- predict(final_gbm,newdata =df_housing_test_cat,n.trees =
10000,type="link")
df_housing_test_cat$SalePrice=Predict_price
df_final<-subset(df_housing_test_cat,select = c(Id,SalePrice))
colnames(df_final)

## [1] "Id"        "SalePrice"

write.csv(df_final,file = "Submission.csv")
```