Predicting Affordability of Houses in Ames, Iowa

STATS 101C - LEC 2: KAT

Albert Na, Kathy Fu, Tiffaney Pi

Overview

- Data Cleaning
  - Type Conversion
  - Handling NAs - Imputation, Zero, ‘None’
  - Correlation
  - Creation + Deletion of Variables
- Methodology
- Results/Interpretation

The data frame

- Clean training and testing datasets together in a new data frame `alldata`
  - `rbind()`
- Remove the 2 observations with an NA value in the affordability column

Type Conversion

- Use mutate_if to save any variables with character types into factors if they weren't already
- Convert categorical variables with a numeric class type into factors based on their descriptions
  - `MSSubClass, OverallCond, OverallQual, BedroomAbvGr, KitchenAbvGr, TotRmsAbvGrd Fireplaces, GarageCars, MoSold, YrSold`

NA values - Imputation

- **Median** → numerical variables (low variance)
  - `LotFrontage`
- **Mode** → categorical variables
  - `MSZoning, Exterior1st/2nd, Electrical, KitchenQual, Functional, SaleType,`
- **Zero** → If the variable had an NA, reasonably assumed this meant the variable did not have the observation at all
  - `MasVnrArea, BsmtFinSF1/2, BsmtUnfSF, TotalBsmtSF, BsmtFullBath, BsmtHalfBath, GarageArea, GarageCars`
  - Ex: If NA value for `MasVnrArea,` assumed that there was no Masonry Veneer Type (level was 'None') to begin with.

NA values → Flagging as "None"

- If NA means 'None', **create factor level 'None'.**
  - i.e.: `PoolQC`: NA means 'No pool'

- `Alley, MasVnrType, BsmtQual, BsmtCond, BsmtExposure, BsmtFinType1/2, FireplaceQu, GarageType, GarageYrBlt, GarageFinish, GarageQual, GarageCond, PoolQC, Fence, MiscFeature`

Correlation

- Check for highly **correlated** variables (r > 0.8 or r < -0.8)
  - Remove `TotRmsAbvGrd`, correlated with `GrLivArea`
  - Remove `GarageCars`, correlated with `GarageArea`
  - Remove `GarageYrBlt`, correlated with `YearBuilt`

*Keep variables that are numerical

Interpreting the Data:

Combining + Creating New Variables

- Create a new variable **`AgeofHouse`**
  - `AgeofHouse = YearRemodAdd - YrBuilt`
- Create new variable **`BsmtBath`**
  - `BsmtBath = BsmtFullBath + .5*BsmtHalfBath`
- Create new variable **`Bath`**
  - `Bath = FullBath + .5*HalfBath`

*Remove old variables and convert new variables into factor type

Interpreting the Data:

Deleting Repetitive Variables

`library(caret)`

- ❖ Apply function `nearZeroVar()` on `testing` and `training`
- ❖ Removed all 24 variables:
  `Street, Alley, LandContour, Utilities, LandSlope, Condition2, RoofMatl, MasVnrArea, BsmtCond, BsmtFinType2, BsmtFinSF2, Heating, LowQualFinSF, KitchenAbvGr, Functional, WoodDeckSF, OpenPorchSF, EnclosedPorch, ThreeSsnPorch, ScreenPorch, PoolArea, PoolQC, MiscFeature, MiscVal`

Interpreting the Data:

Deleting Other Variables

- Remove numerical variables:
  - `Obs` - not actually a predictor
  - Use `geom_density()` on remaining numerical variables
    - Remove `LotFrontage` and `LotArea`
- Remove categorical variables:
  - `Neighborhood` - too many levels
  - Table these variables, remove ones with infrequent levels of around <1000 occurrences
    - Remove `BldgType, RoofStyle, CentralAir, Electrical, GarageQual, GarageCond, PavedDrive, Fence`

Attempting different Methods

1) ~~Logistic Regression~~
2) ~~Lasso, Ridge Regression~~
3) ~~SVM~~
4) ~~Xgboost~~
5) ~~Tree~~

Random Forest

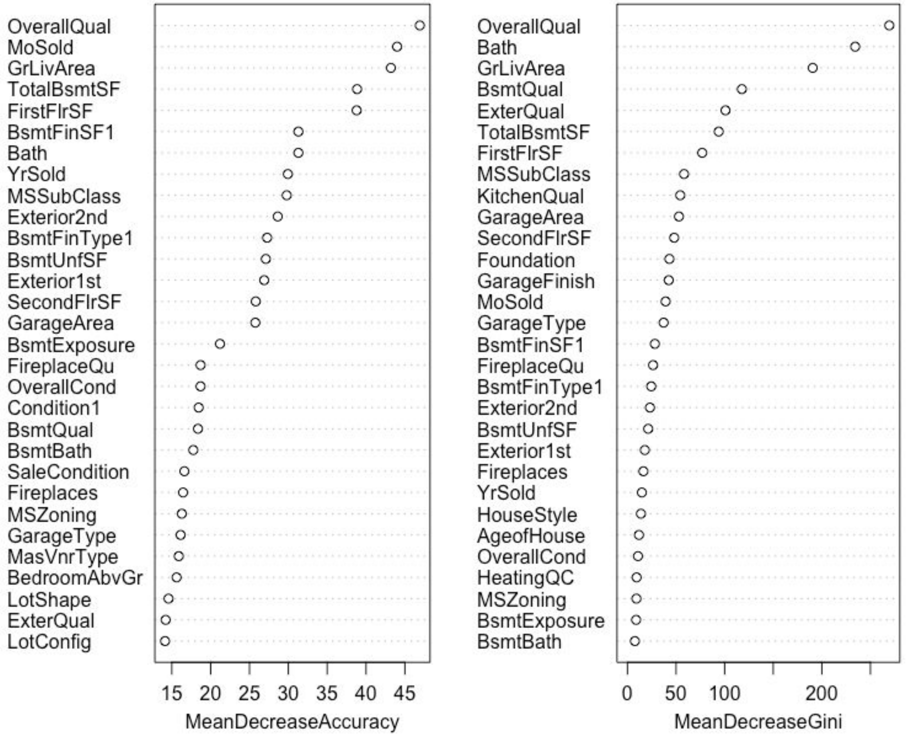
Random Forest

- Final dimensions of training data: **3498 x 39** (Originally 3500 x 81)
- mtry=9

Confusion Matrix

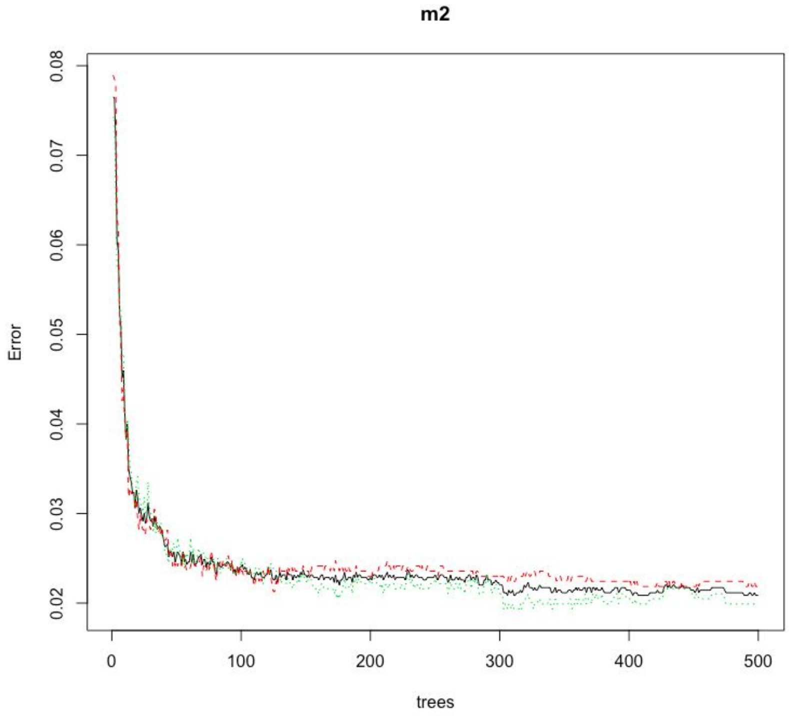
	Affordable	Unaffordable
Affordable	1701	38
Unaffordable	35	1724

Variable Importance



- ❖ Most significant predictors at the top.
- ❖ Model contained all predictors

Classification error



- ❖ Sufficient number of trees ~ 100
- ❖ **OOB** = "Out of Bag" error ~ 2.09%
  - ➢ running unbiased estimate of the classification error as trees are added to the **forest**
- ❖ "Affordable" error rate: 2.185%
- ❖ "Unaffordable" error rate: 1.99%

Final Accuracy: 98.89%

Private leaderboard: 97.90%

Limitations & Recommendations

- Public leaderboard not representative of private leaderboard
- Overfitting
- Sample size
- With more time, perform deeper analysis on each variable and consider using outside data. Possibly combine infrequent levels together into “Other” category.

Thank you!

Figure 2:

Figure 3:

Figure 1:

Figure 4:

Figure 6:

Figure 5:

Figure 7:

Figure 8:

	Affordable	Unaffordable
Affordable	1705	34
Unaffordable	40	1719