The Dummy Variables

Valerie Huang xiaoyinghuang@ucla.edu
Konner Macias konnermacias@ucla.edu
Breanna Ramos breanna.ramos25@yahoo.com
Hayley Todd hayleytodd27@gmail.com

Background

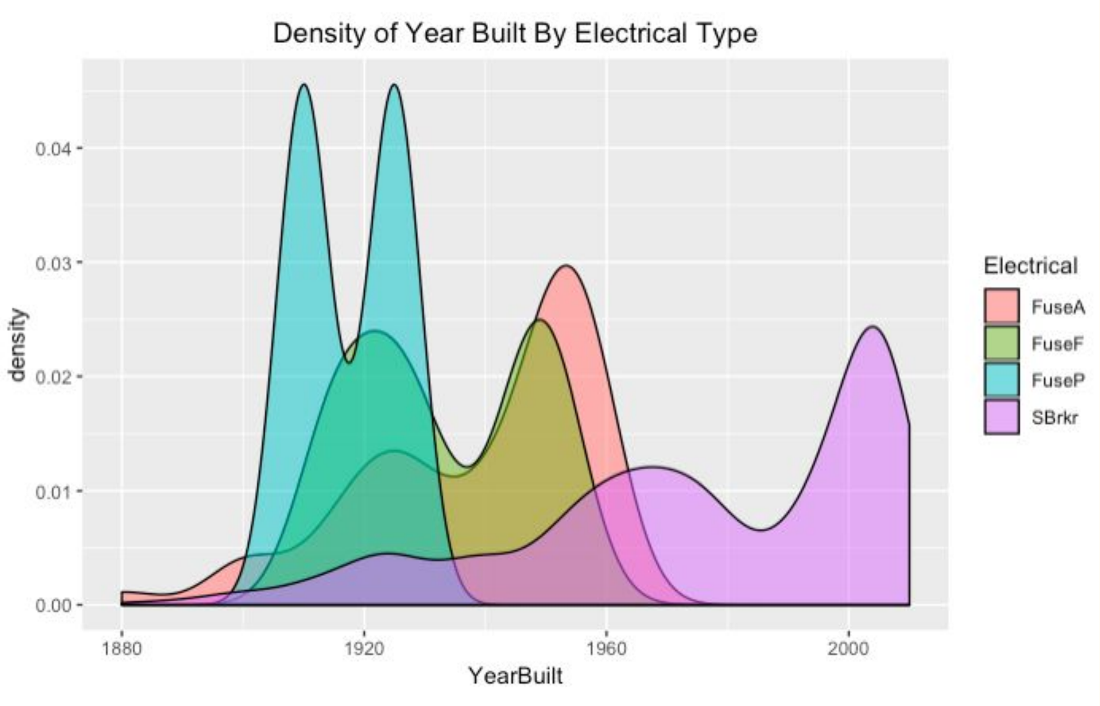
- Dataset: **Ames Housing** data compiled by Dean De Cock
  - **3500** observations
  - **79** descriptive variables
- Project Question:
  - Using the above data, **can we classify** whether a given house is affordable or not?
- Our Goals:
  - Investigate each variable, understand relationships, clean, and check **whether new variables can be created**
  - Compare across industry standard classification techniques, and tune an appropriate model for classification.

Cleaning the Data

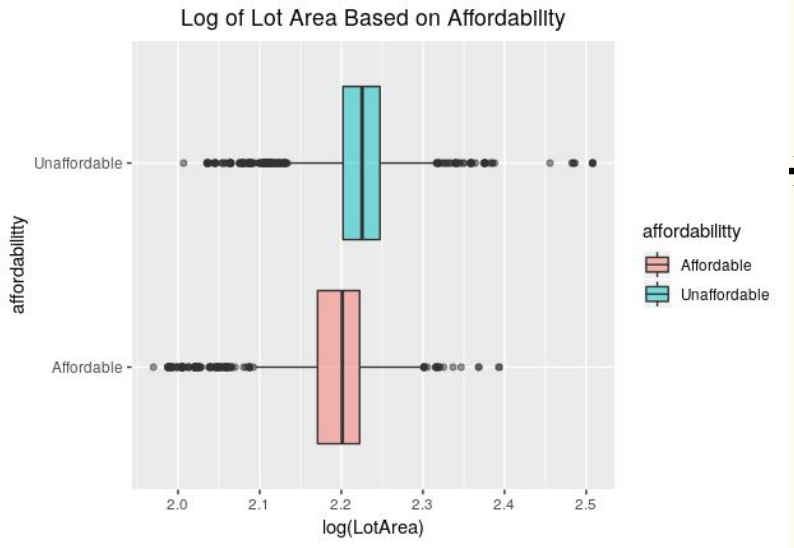
- ➔ Within the raw data, there were **32 variables that had NAs.**
  - ◆ For 15 out of those 32 variables, **NA represented "None"**.
- ➔ Still 17 variables, such as LotFrontage, MSZoning, and MasVnrType that contained missing values → LotFrontage had 560 → Eliminated
- ➔ Variables that had integers representing categories (such as MSSubclass and Quality/Condition variables) were **changed to factors**.
- ➔ After some manual observations:
  - ◆ Many NAs for variables (primarily Basement of Garage variables) had None in their related columns → **Changed to None as well**
- ➔ While some variables had obvious changes, others required **a little more thought and effort.**

Cleaning the Data

- For variables with very few missing values (such as Utilities and Electrical), we looked at those observations manually and **chose the appropriate value based on other key variables.**
- Variables with a higher amount of NAs were filled in by imputing with **mice.**



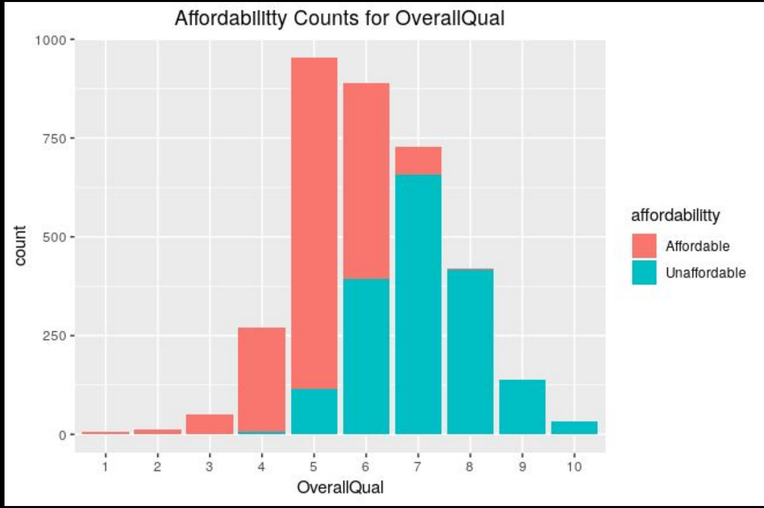
Process for Variable Exploration

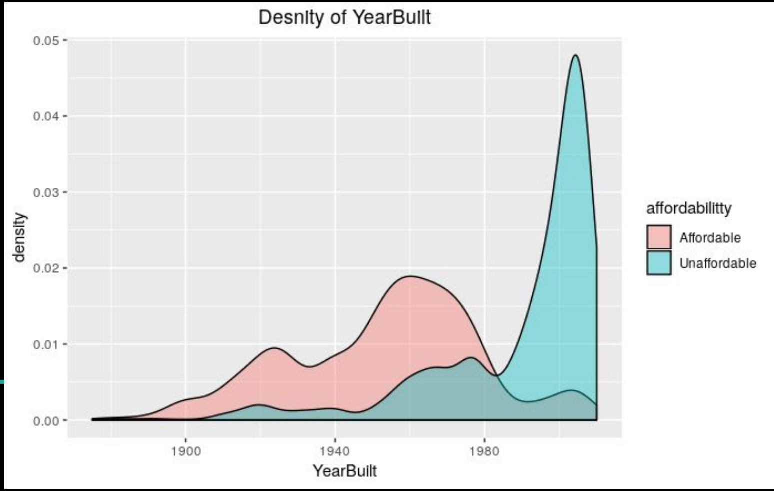


- ➔ Once we had cleaned most of the data that originally came with the data set, we moved on to variable exploration for future selection.
- ➔ For each individual variable:
  - ◆ Plot vs Affordability
  - ◆ Run basic glm models
    - Check misclassification
  - ◆ Group similar variables
    - Run more glm models
    - Check for VIF multicollinearity

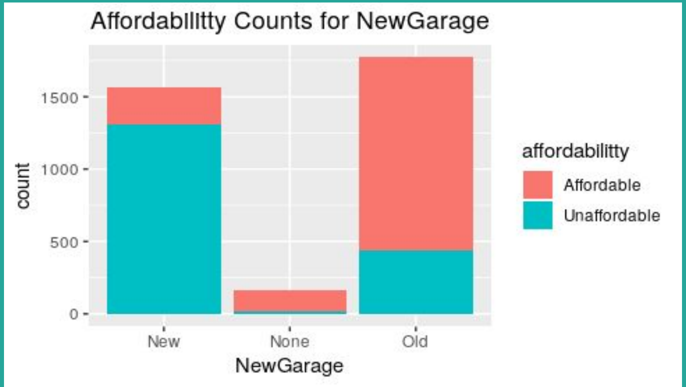
Determining Key Variables

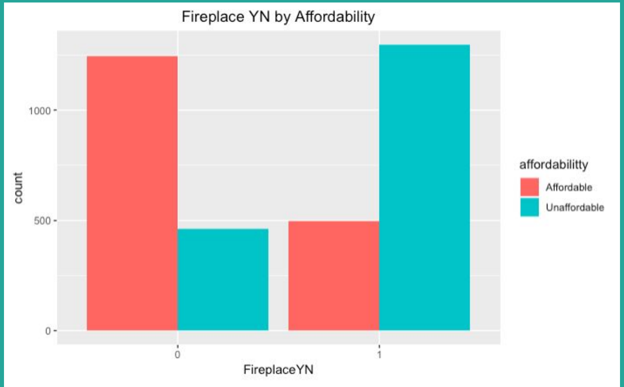
- ➔ By plotting the variables and running individual/group linear model misclassification tests, we were able to determine which were advantageous
- ➔ We were able to initially speculate that Neighborhood, OverallQual, and YearBuilt would be strong predictors.





Creating New Variables





- ➔ There were several variables such as OpenPorchSF, X3SsnPorch, EnclosedPorch, ScreenPorch that were not as informative on their own.
- ➔ Therefore, we created a binary variable (PorchYN) that was 0 if the house lacked a porch, and 1 if the house has a porch
- ➔ We also created:
  - ◆ TotBath
  - ◆ FireplaceYN
  - ◆ HasMsn
  - ◆ FenceYN
  - ◆ NewGarage

Our Resulting Model

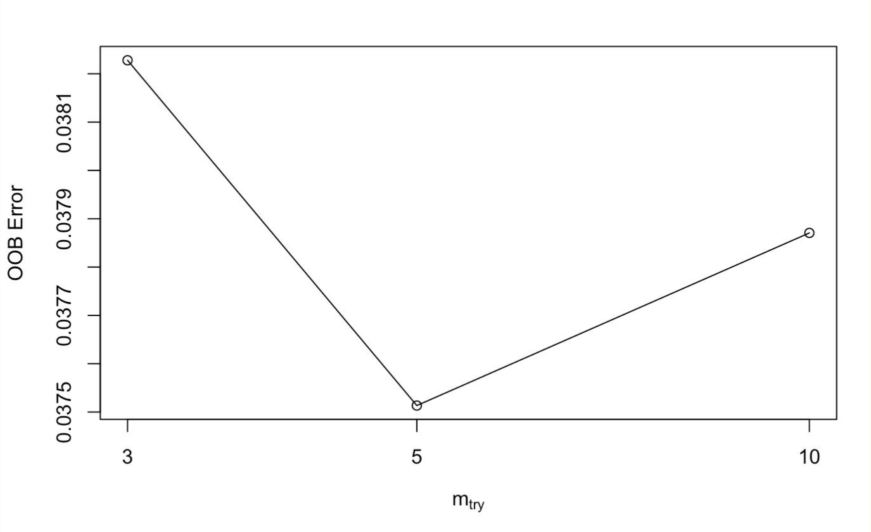
- ➔ After all of our tests and observations we chose to **eliminate 59 variables** from our model, while **adding 6 new variables** of our creation.
- ➔ Our final model contained 27 variables:
  - ◆ MSSubClass, MSZoning, LotArea, LotShape
  - ◆ Neighborhood, OverallQual, YearBuilt,ExterQual
  - ◆ Foundation, BsmtQual, TotalBsmtSF, HeatingQC
  - ◆ GrLivArea, BedroomAbvGr, KitchenQual, TotRmsAbvGrd
  - ◆ GarageType, GarageFinish, GarageCars, GarageArea
  - ◆ WoodDeckSF, MoSold, PorchYN, FireplaceYN, HasMsn,
  - ◆ TotBath, FenceYN, NewGarage

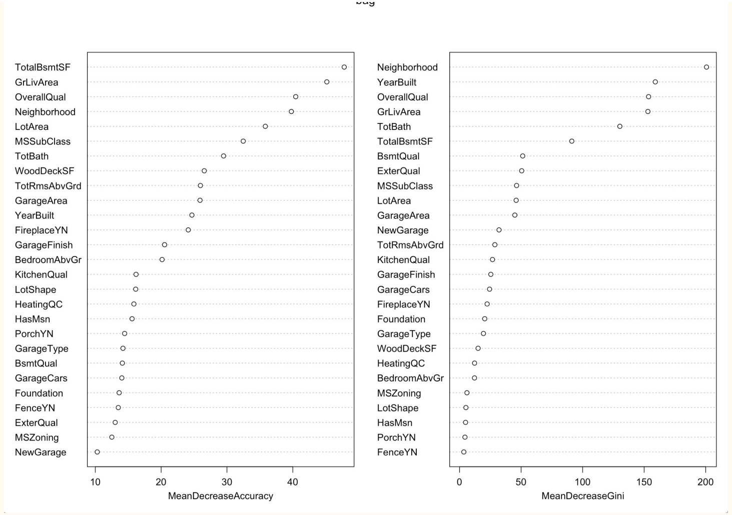
Cleaning the Testing Data

- Our final step before modelling and making predictions was cleaning the testing data.
  - First, it was necessary to do **similar cleaning** and transformations to that of the training data such as converting the "NA"s to "None" for the appropriate variables, while **also imputing** certain missing values for other variables.
  - We then needed to add our newly created variables to our testing data.
  - We also discovered that many variables within the testing data **had a different number of levels of factors** than the training data, so we had to convert those testing levels as well.

Methodology

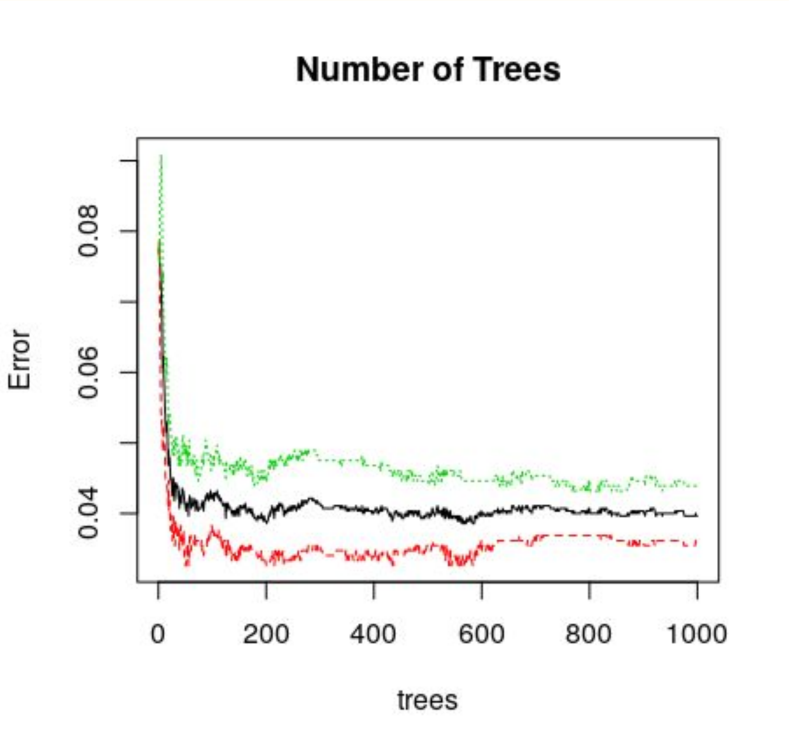
- ➔ We split the training data into 80% training and 20% testing.
- ➔ After trying out logistic regression, various tree models, SVM, etc., we chose random forest for our final model.
- ➔ The Variable Importance Plot also showed that most of the variables we picked were useful in predicting affordability





Model Performance

- ➔ Resulting Model:
  - Full Random Forest
  - Mtry: 5
  - Number of trees: 500
- ➔ Results on 20% of testing data
  - Misclassification rate: 1.14%
    - (8 out of 699 incorrect cases)



Main Results

- Public Leaderboard
  - On 50% of the test data, we obtained an **98.4% accuracy**
- Private Leaderboard
  - On the remaining 50% of the test data, we obtained **98.0% accuracy**
    - This submission categorized 754 houses as Affordable, and 746 as Unaffordable

Limitations

- Initially, when we assessed the graphs we included 43 variables in our model
- We failed to realize that many of these variables were insignificant
- Despite removing these insignificant variables from our model, we were only able to achieve 98.4% accuracy on the public leaderboard
- We also **included two variables with high multicollinearity** in our model: MSZoning and Neighborhood
- Finally, our Ensemble Method included techniques that were **very similar**

Recommendations

- ➔ Most models we tried were **tree-related models** (tree classification, bagging, random forest,boosting) and **“majority vote” ensemble** models with those tree-related ones. **All generated similar results.**
- ➔ If we have three uncorrelated models that can each explain more than 97, 98% of the training data, we could again try the “majority vote” approach with those instead.
  - ◆ We believe that this method could improve our prediction accuracy.

Conclusions

- Not all variables are useful in predicting affordability. TotalBsmtSF, GrLivArea, OverallQual, Neighborhood, LotArea, MSSubClass, TotBath, GarageArea and FireplaceYN are the **most significant** predictors.
- Despite ranking 37th on the public leaderboard, we **ranked 7th on the private leaderboard**
- This indicates that our model **did not overfit.**
- Further exploration and new techniques should be considered.