## Homework 6- DS450

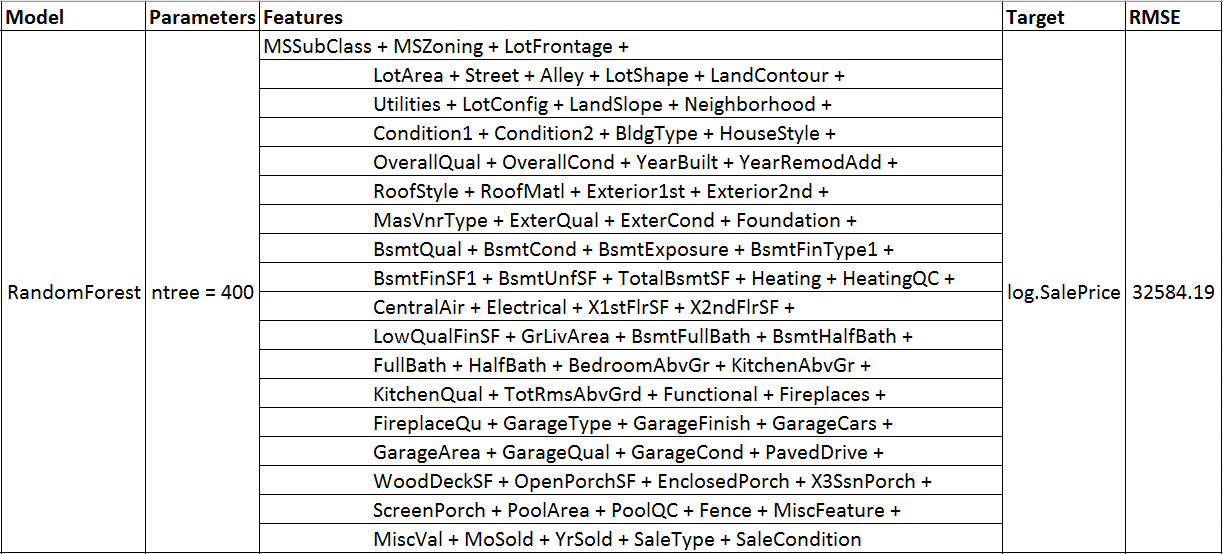
Submitted By: Danny Godbout

## Data Preparation

* Factor columns with NAs are re-labeled “None”
  + Exception: Electrical is replaced with “SBrkr”
* Numeric columns with NAs are re-labeled with 0
  + Exception: LotFrontage is replaced with the mean of all LotFrontage values
* For models 2 & 3, the following factor features are re-binned to handle small categories that may not appear in the training set, or are too small to model. The default cutoff requires a level appear in >2% of records, with exceptions noted in brackets below:
  + Neighborhood
  + MSSubClass
  + Exterior1st
  + SaleType
  + HouseStyle
  + Functional
  + GarageType
  + RoofStyle
  + FireplaceQu
  + GarageQual [5%]
  + SaleCond
  + MSZoning [5%]
  + BsmtExposure
  + LotShape
  + MasVnrType
  + KitchenQual [3%]

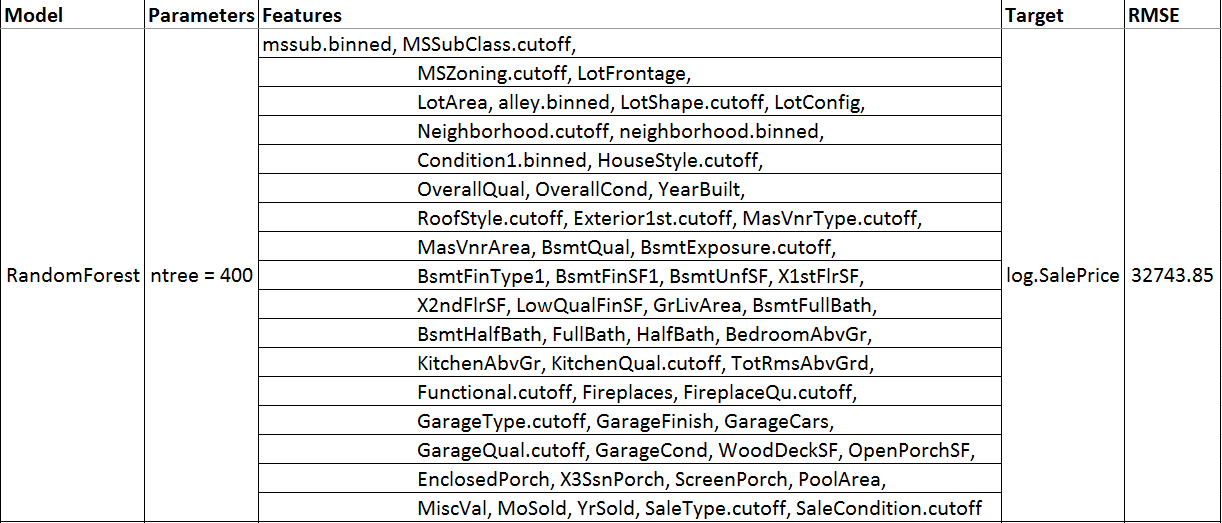
## Model 1- Random Forest on Clean Data

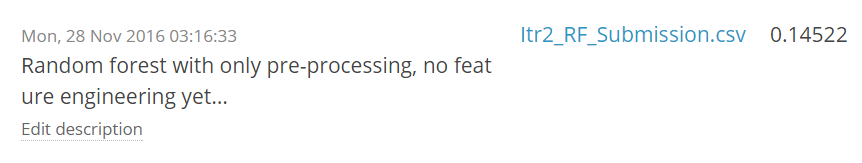
Using only the clean data, before any processing (e.g. discretization, re-leveling factors, etc.).

Unfortunately, this model fails to execute on the competition data due to small factor levels that are not found in the training set. This requires some processing of high-cardinality factors to reduce the number of levels.

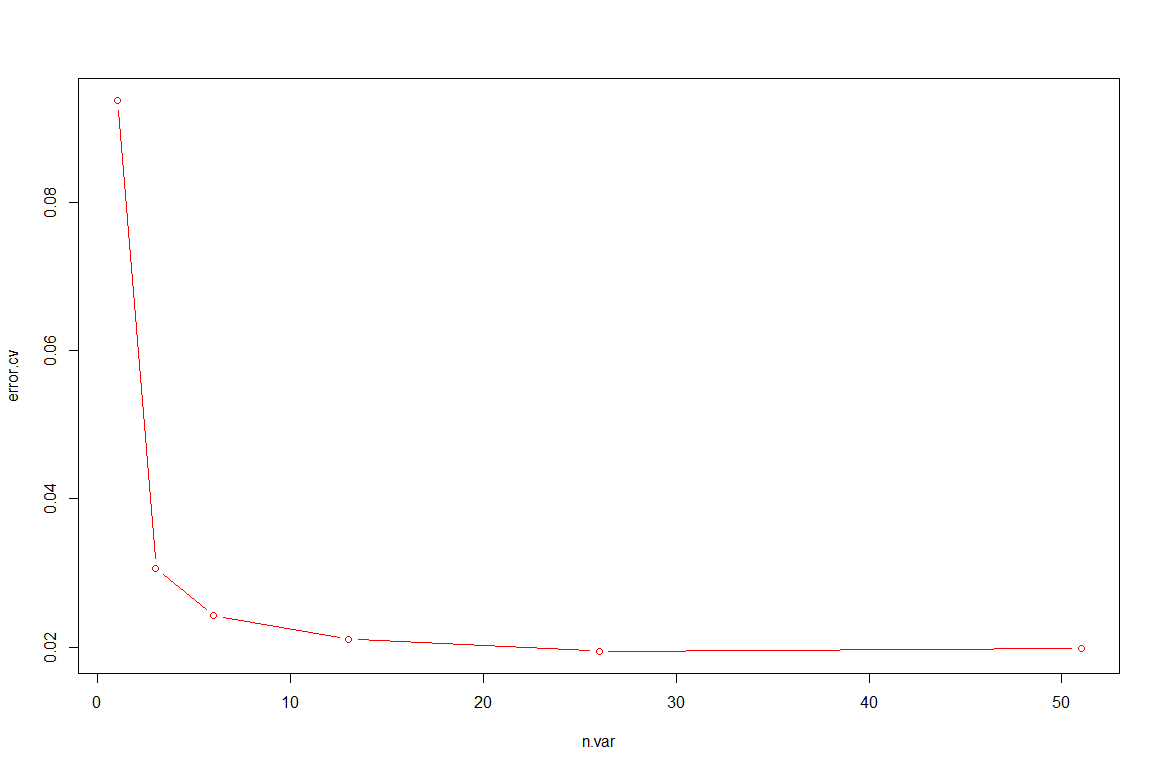
## Model 3- Random Forest on Binned Data

This iteration adds a processing step to the input data to combine factor levels with small numbers of observations into an “Other” level.

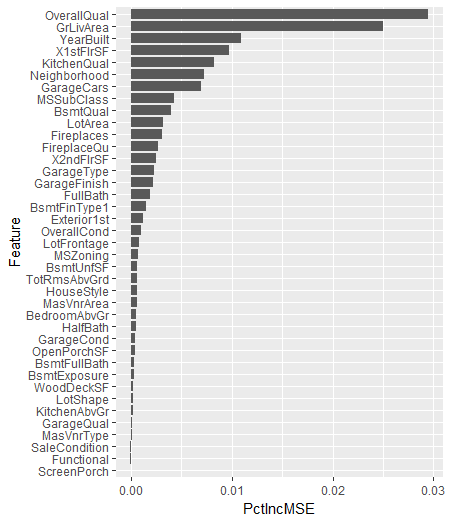
While RMSE nominally decreased, the model is more robust to new and unique factor levels and returned a Kaggle score of 0.14522:



Cross-validation of the model shows that the model error settles when using approximately 25 features, and error increases slightly when using the full feature set.

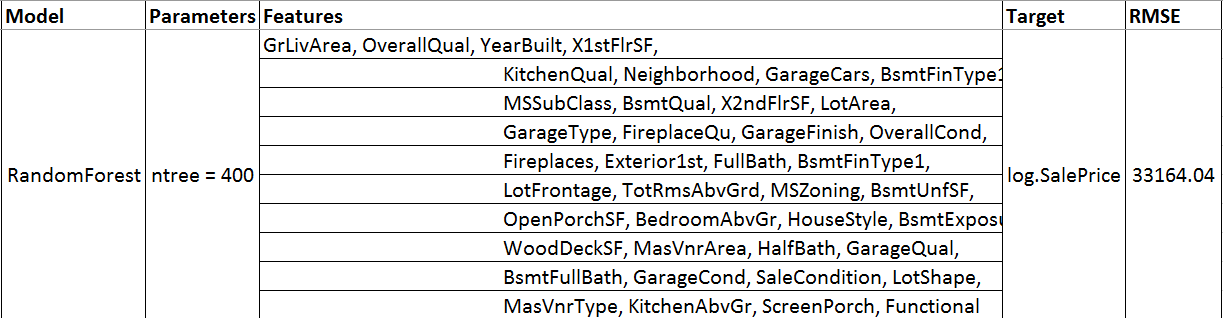


A variable importance plot was generated from the Random Forest to select the top features; the top 39 features are shown below.



## Model 3- Random Forest on Binned Data, with Feature Selection

After selecting the top features, the RandomForest model was re-trained.

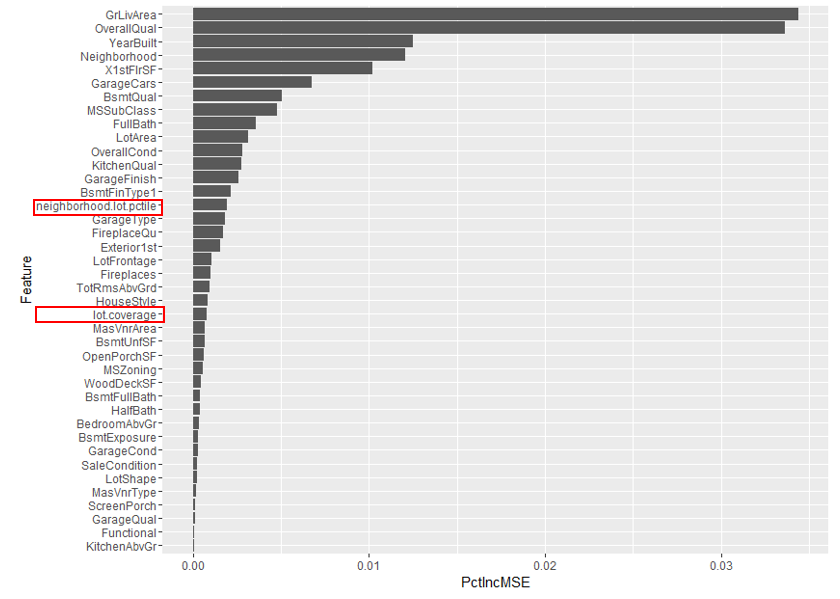


With this iteration, RMSE increased slightly and the Kaggle score failed to surpass the submission for Model 2.



## Model 4- Combine RandomForest, xgBoost, and Linear Model using Stacked Generalization + Feature Engineering

This iteration adds multiple changes:

* Adds xgBoost and Linear Model predictions, then blends all predictions using a top-level Random Forest learner
* Performs feature log transformations on the following:
  + GrLivArea
  + X1stFlrSF
  + LotArea
  + TotRmsAbvGrd
* New engineered features are added:
  + neighborhood.lot.pctile: Divide LotArea by the mean of the Neighborhood LotArea to get a sense of “how big is this lot for this neighborhood?”
  + lot.coverage: Captures big house on small lot scenarios by dividing X1stFlrSF (home outline) by LotArea
  + The new features performed well on feature importance:  
    

The combination of the ensemble approach and new engineered features produced an RMSE improvement:



The improved RMSE on the test split translated to significant performance improvements in the Kaggle ranking:

