KAGGLE ML PROJECT

House Prices: Advanced Regression Techniques

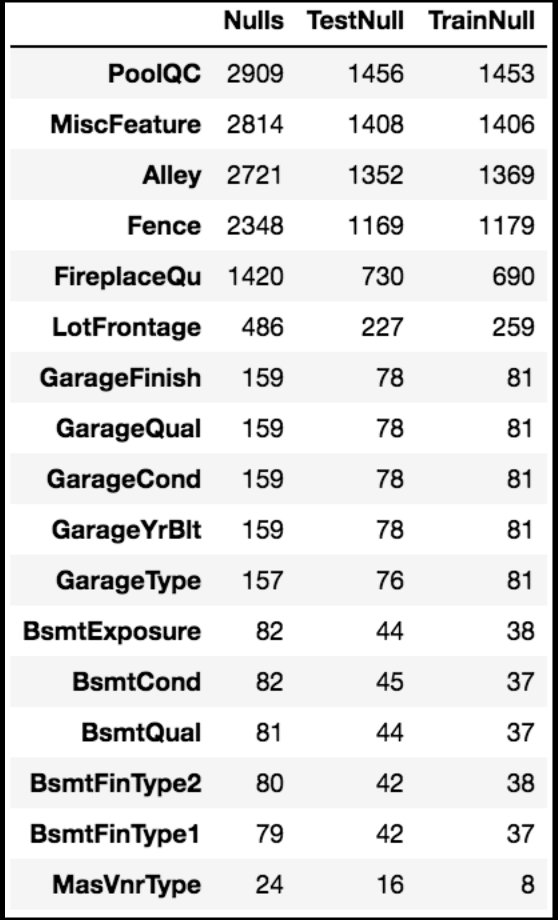
Team FightingMongooses: Billy Fallon, Gregory Brucchieri, Adrian Phillips-Samuels

INTRODUCTION

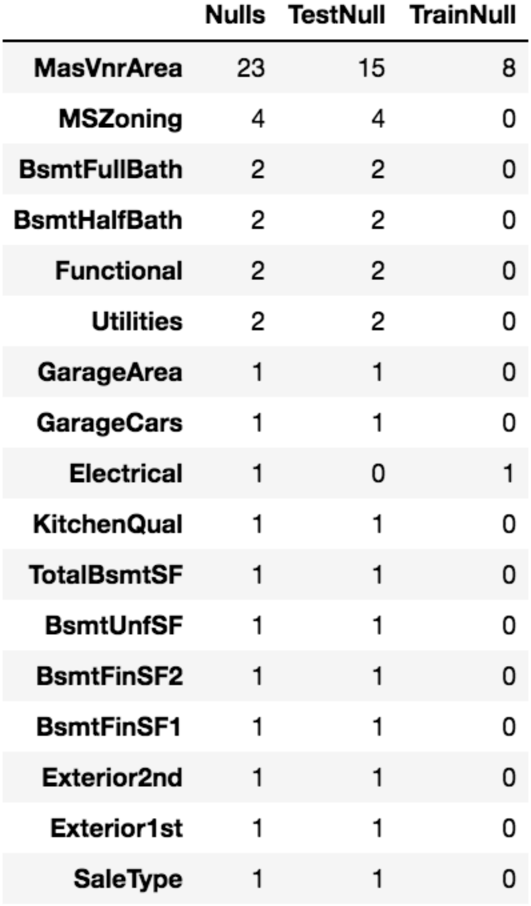
- Data:
  - 79 Features
  - Target variable: SalePrice
  - 1460 observations in training set, 1459 in test set
  - Mix of numerical, categorical, ordinal
    - 28 continuous
    - 51 categorical/ordinal

MISSINGNESS

- 34 features missing some data
- We used various techniques for handling missingness



	Nulls	TestNull	TrainNull
PoolQC	2909	1456	1453
MiscFeature	2814	1408	1406
Alley	2721	1352	1369
Fence	2348	1169	1179
FireplaceQu	1420	730	690
LotFrontage	486	227	259
GarageFinish	159	78	81
GarageQual	159	78	81
GarageCond	159	78	81
GarageYrBlt	159	78	81
GarageType	157	76	81
BsmtExposure	82	44	38
BsmtCond	82	45	37
BsmtQual	81	44	37
BsmtFinType2	80	42	38
BsmtFinType1	79	42	37
MasVnrType	24	16	8



	Nulls	TestNull	TrainNull
MasVnrArea	23	15	8
MSZoning	4	4	0
BsmtFullBath	2	2	0
BsmtHalfBath	2	2	0
Functional	2	2	0
Utilities	2	2	0
GarageArea	1	1	0
GarageCars	1	1	0
Electrical	1	0	1
KitchenQual	1	1	0
TotalBsmtSF	1	1	0
BsmtUnfSF	1	1	0
BsmtFinSF2	1	1	0
BsmtFinSF1	1	1	0
Exterior2nd	1	1	0
Exterior1st	1	1	0
SaleType	1	1	0

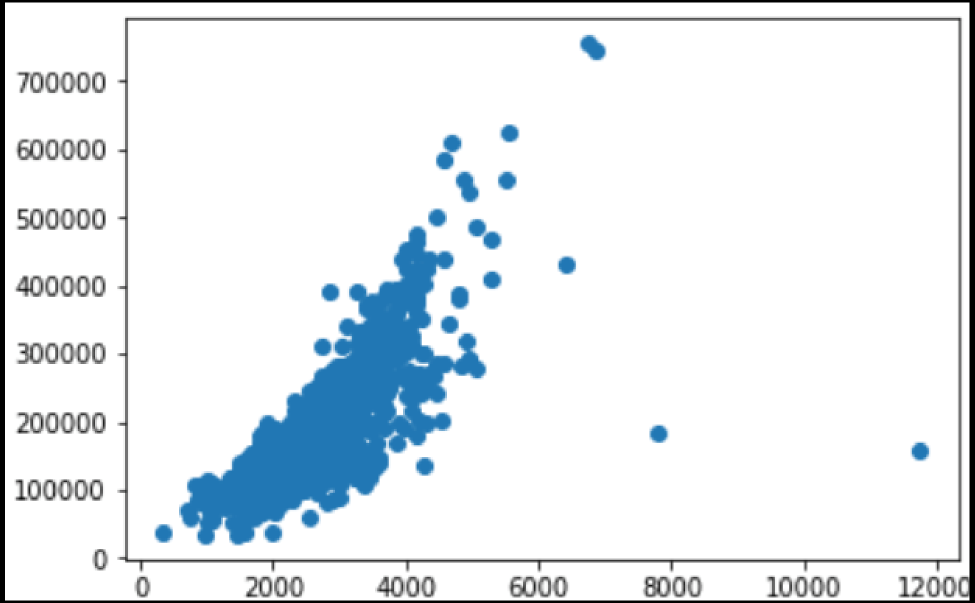
HANDLING MISSINGNESS

- Garage Year Built = (oldest house year -1)
- Electrical = “unknown”
- Zone, exterior, sale type: missing values in the test set: imputed by comparison to similar properties in neighborhood
- LotFrontage = mean
- Others: 0 or None as appropriate

PROCESSING DATA

- Treated features as ordinal whenever possible:
  - PavedDrive, GarageFinish, Functional, BsmtFinishType1,2, BsmtExposure, LandSlope, Utilities, LotShape, Alley, Street, CentralAir

- Observed two outliers
- Unusual Price/TotalSF ratio
- Chose Robust Scaling for this reason



ENGINEERING FEATURES

- Calculated TotalSF
- Calculated PorchSF
- Converted Porch type to dummies
- Replaced remodel year with a Boolean value “Remodeled”
- ‘Normalization – used Robust Scalar
- Transformation: Log(SalePrice), Log(LotArea)
- Tried:
  - Recession dummy – 1 if Sale Dec2007-June2009, else 0
  - Total Baths
  - Years old vs. Year build

MODELS USED

- Linear Regression
- Random Forest
- Gradient Boosting Regression
- XGBoost

MODEL 1 – RANDOM FOREST

- Parameters: bootstrap=False, max_features='sqrt', min_samples_leaf=1, min_samples_split=2, n_estimators=800
- Top Features:
  - TotalSF
  - Overall Quality
  - Gross Living Area
  - Exterior Quality
- RMSE: 0.0192
- KAGGLE: 0.14294

MODEL 2 – GRADIENT BOOST

- Parameters: alpha=0.9, learning_rate=0.01, max_depth=5,
  max_features='sqrt', min_samples_leaf=1,
  min_samples_split=10, n_estimators=1200
- Top Features:
  - Overall Condition
  - Total SF
  - Lot Area
  - Gross Living Area
- RMSE: 0.0148
- KAGGLE: 0.12707

MODEL 3 – XGBOOST

- Parameters: base_score=0.5, booster='gbtree', gamma=0.3, learning_rate=0.1, max_depth=3, n_estimators=800, subsample=0.6
- Top Features:
  - Total SF
  - Overall Condition
  - Lot Area
  - Overall Quality
- RMSE: 0.0179
- KAGGLE: 0.13783

MODEL4 – LINEAR REGRESSION

- Parameters: Nothing Interesting

- Top Features:
  - That would be neat.
  - First ensemble attempt combined Linear Regression and Random Forest (2:1) and yielded a 0.1295

- RMSE: 0.0353
- KAGGLE: 0.13691

ENSEMBLING

- Simple Average improved upon all single model results 0.12671
- Stacked Regression produced impressive CV scores, and AWFUL Kaggle Scores
  - Severe Overfit?
  - Implementation Failure?
- Weighted Averages improved upon the Kaggle Score
- 60-30-10 weighting of top three models (GBM, Regression, XGB) produced best Kaggle Score: 0.12320

CONCLUSIONS/NEXT STEPS

- Feature Engineering allows for infinite permutations, and will make or break the outcome
- Would like to experiment further with other scalars with this data sets, and neighborhood specific modelling
- Iterative Kaggle submission is addicting – should come with a food pellet