Kaggle Housing Data

NYC Data Science Academy
Fall 2018 Cohort

Tristan Dresbach
Karim Zaatary
Sean Justice

Missingness

- 19 Features had missing values
  - Missing Completely at Random: 3
  - Missing at Random: 16
- Mostly related to lack of feature
- Continuous values were imputed by mean
- Categorical values were imputed by random selection

Initial Features Removed

Id	Not related to Price
Exterior2nd	Highly correlated with Exterior1st
TotRmsAbvGrd	Highly correlated with GrLivArea and BedroomAbvGr
GarageArea	Highly correlated with GarageCars
BsmtFin1SF	Sums equal to TotalBsmtSF
BsmtFin2SF	
BsmtUnfSF	
1stFlrSF	Sums equal to GrLivArea
2ndFlrSF	

Types of Feature Selection

- Univariate
- Step forward
- Step Back

Type	Categorical features	Continuous features	Total
Step forward	52	9	61
Step back	53	14	67
Chi-squared & F-reg	6	21	27
Anova/t-test & F-reg	23	21	44

Key Features

Rank	Step Forward	Step Back	Chi-squared F_reg	Anova/t-test F_reg	Total Data
1	GrLivArea	Condition2_PosN	ExterCond_Ex	Functional_Maj2	RoofMatl_ClyTile
2	YearBuilt	OverallQual_1	ExterCond_Po	RoofMatl_WdShngl	Condition2_PosN
3	GarageCars	MSZoning_C (all)	MiscFeature_TenC	CentralAir_N	RoofMatl_Membran
4	TotalBsmtSF	Functional_Maj2	Street_Pave	ExterCond_Fa	PoolQC_Not_Avail
5	RoofMatl_ClyTile	OverallQual_10	KitchenAbvGr	SaleType_New	RoofMatl_Metal

Univariate Feature Selection

	Ridge with CV		Lasso with CV		XGBoost		Random Forest	
Score	R squared	RMSE	R squared	RMSE	R squared	RMSE	R squared	RMSE
Full data set	74%	0.198	85%	0.185	88%	0.134	88%	0.135
Chi_squared & F_reg	67%	0.225	81%	0.214	85%	0.152	84%	0.156
Anova/t-test & F_reg	72%	0.207	82%	0.209	87%	0.140	86%	0.143

Step Forward

- Started with 281 features (dummified categorical and ordinal variables)
- Used Stepwise feature selection
- Minimized BIC at each step
- Resulted in 61 variables
- 9 continuous
- 52 dummified categorical

Stepwise Forward

	Linear Regression		Ridge with CV		Lasso with CV		Random Forest	
Score	R squared	RMSE	R squared	RMSE	R squared	RMSE	R squared	RMSE
Test	69.2%	0.217	80.5%	0.172	69.2%	0.216	88.95%	0.13

Step Back

- Started with Full Model
  - 281 features after dummifiying
- Ran linear regression
  - Reduced features based on p-value
- Found minimum BIC
  - Finished with 67 features

Step Back Results

	Multilinear		RidgeCV		Lasso CV		Random Forest		xGBoost		Multilinear PCA	
Score	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2
Test	0.201	71%	0.163	82%	0.159	83%	0.137	84%	0.129	89%	0.142	87%

Conclusion