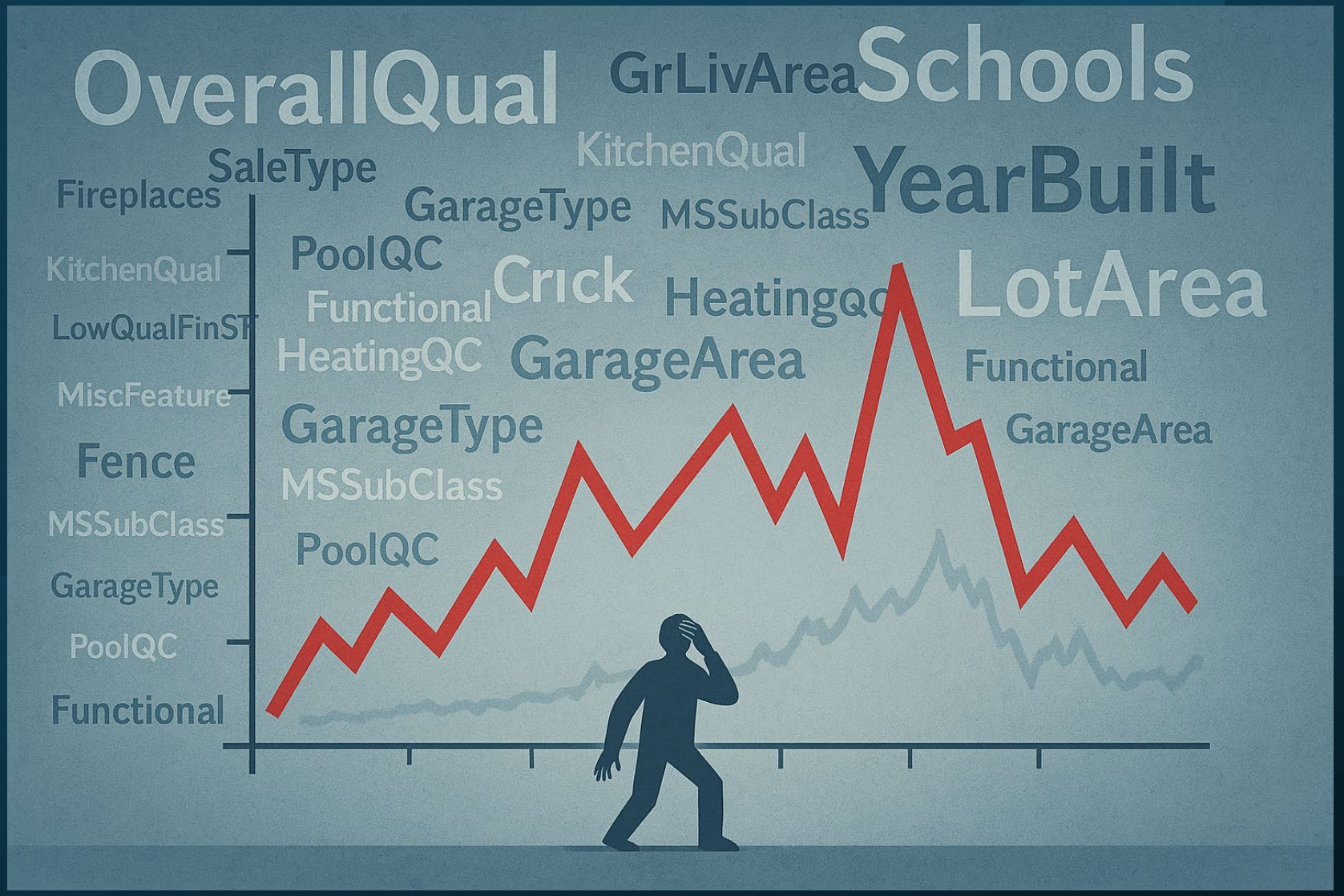
Capstone Project: Predicting House Price in Ames, Iowa

Sheng Miao.PhD

The Issue

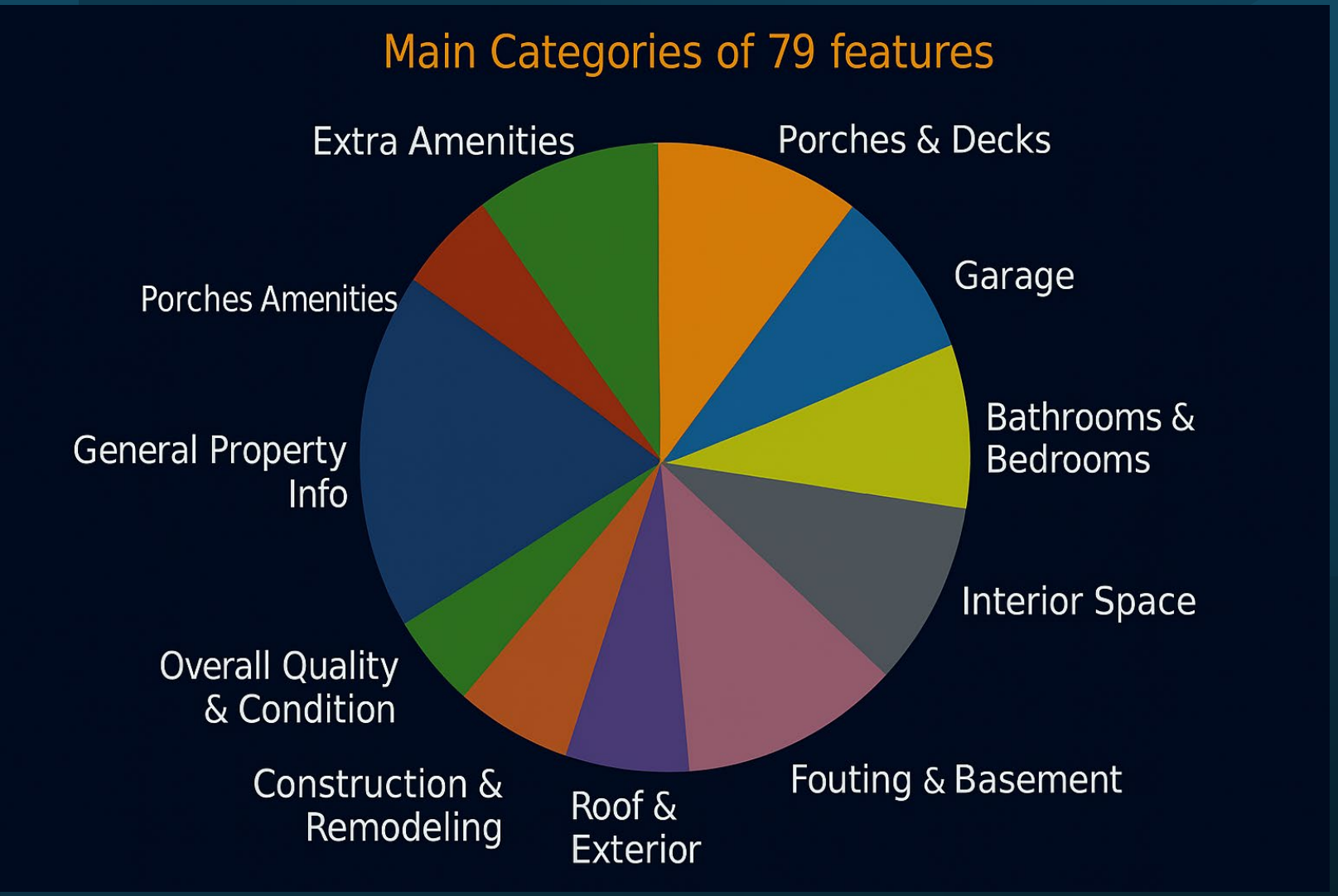

OverallQual
GrLivArea
Schools
SaleType
KitchenQual
YearBuilt
Fireplaces
GarageType
MSSubClass
PoolQC
KitchenQual
Crick
LotArea
Functional
HeatingQC
GarageArea
Functional
GarageType
GarageArea
Fence
MSSubClass
PoolQC
MSSubClass
GarageType
PoolQC
Functional

Who might care?

- Insurance Companies
- Real estate professionals
- Homebuyers
- Sellers
- Real estate Investors
- Data Scientists & Analysts
- Construction & Development Firms
- Banks & Mortgage Lenders
- Retirees & Estate Planners
- Policy Makers & Economists
- Urban Planners & City Officials
- Academic Researchers

What factors might affect house price?

Data Information



Data: Ames Housing dataset

Source: Kaggle

Location: residential homes in Ames, Iowa

Year Sold: 2006-2010

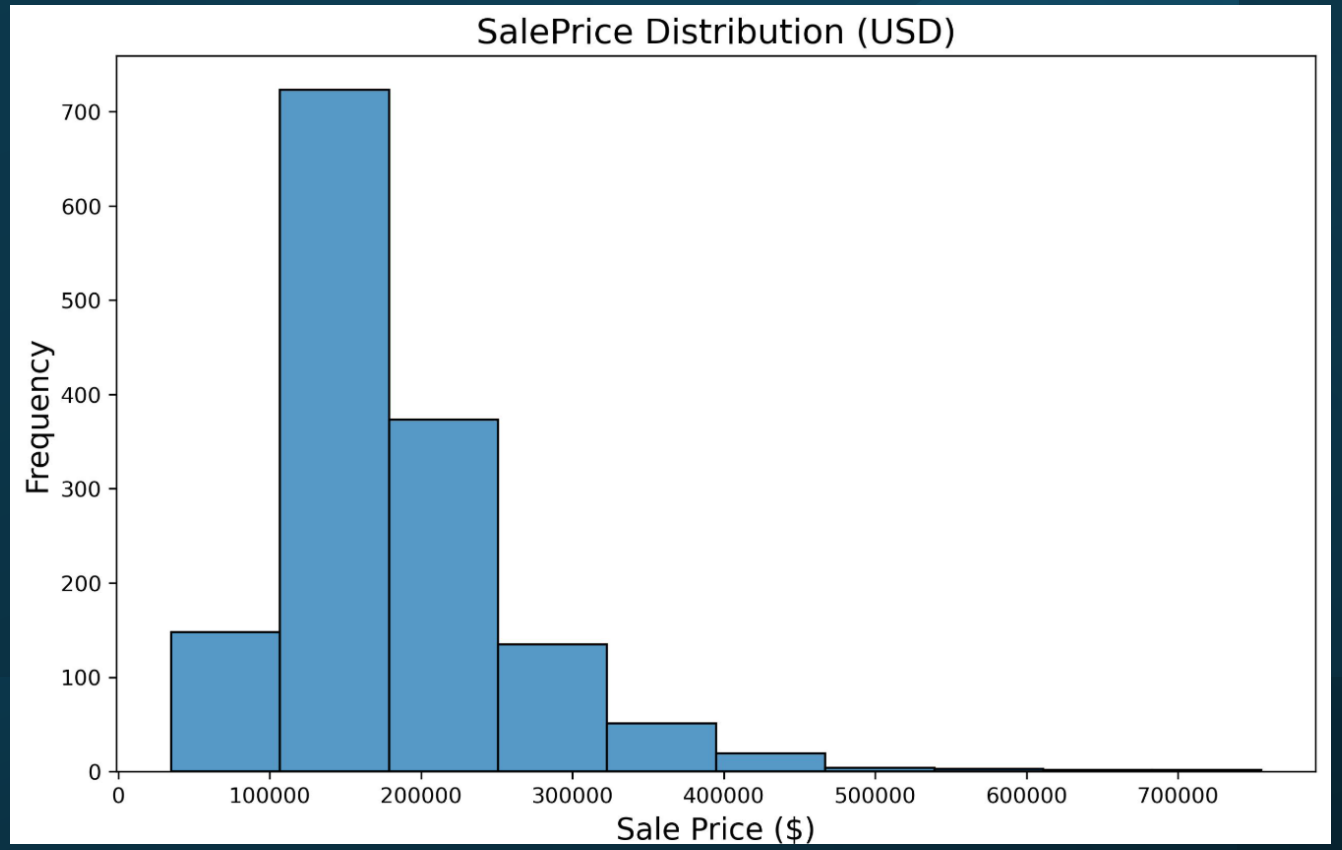
Data size:
Training data -- 1460 records * 79 features,
Testing data -- 1460 records * 78 features,

Data Cleaning and wrangling

- **Missing value cleaning**
- **Rare category combination**
- **Removing features with highly imbalanced distributions and limited predictive relevance**
- **Categorical feature convertion**

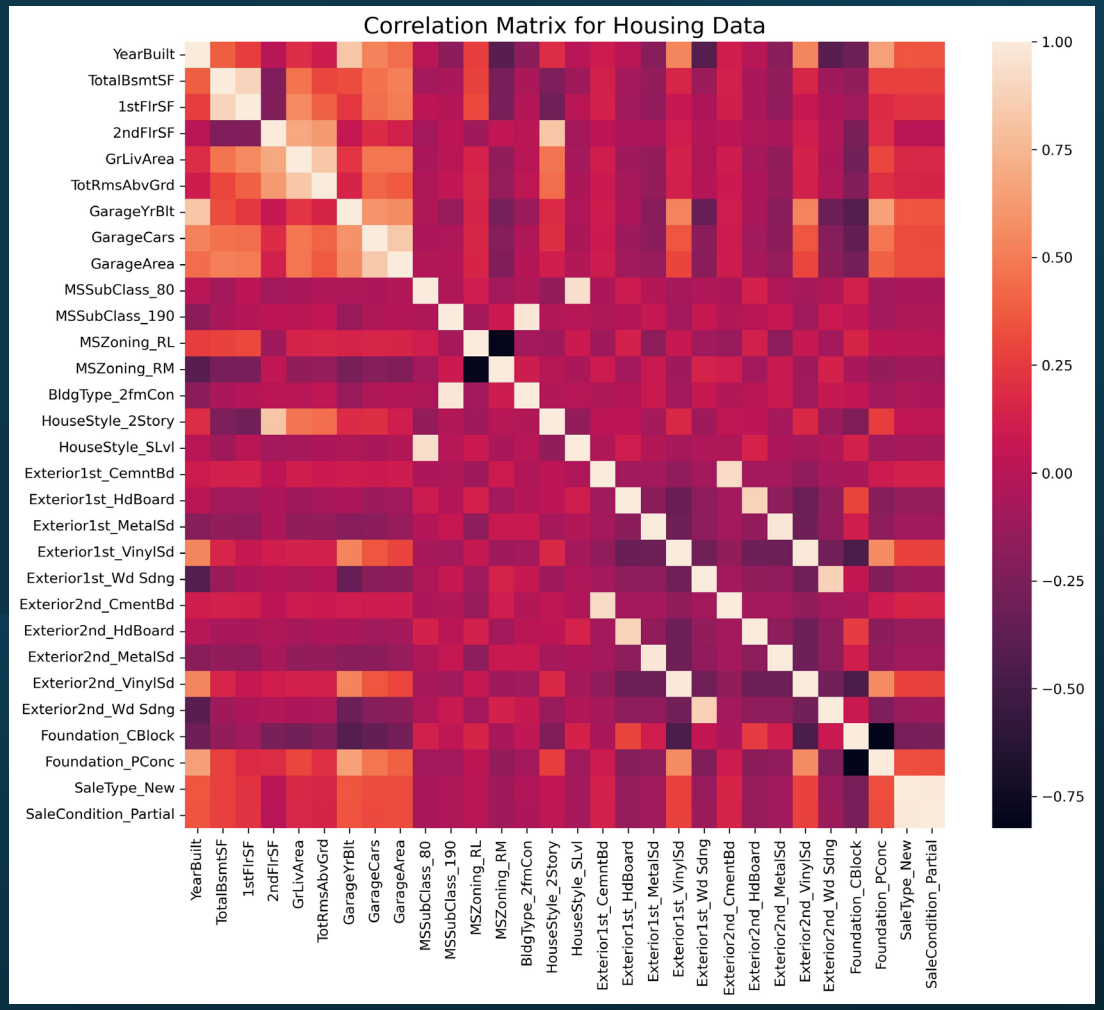
Data Exploration

Data Exploration

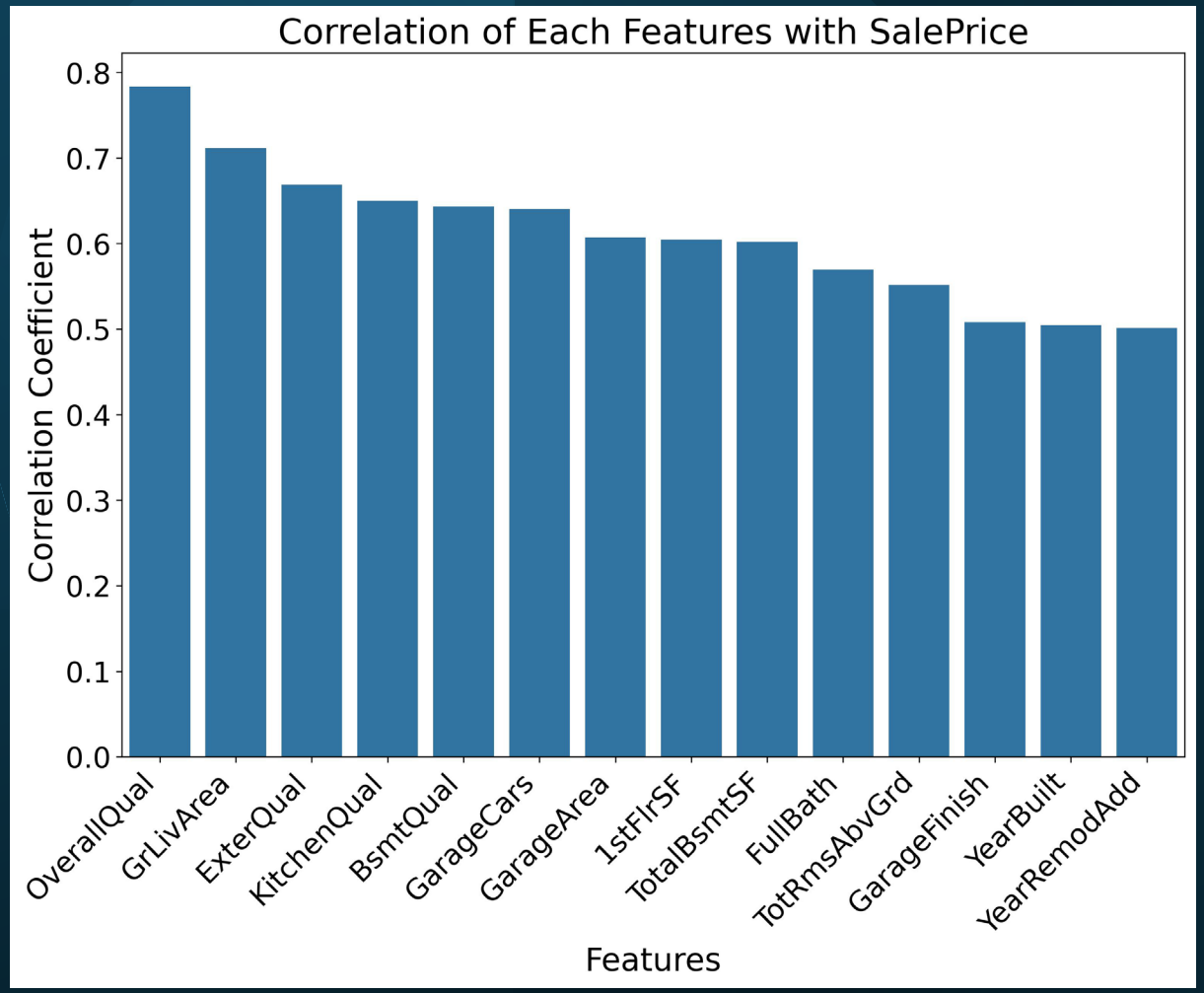


Data Exploration

Correlation Matrix

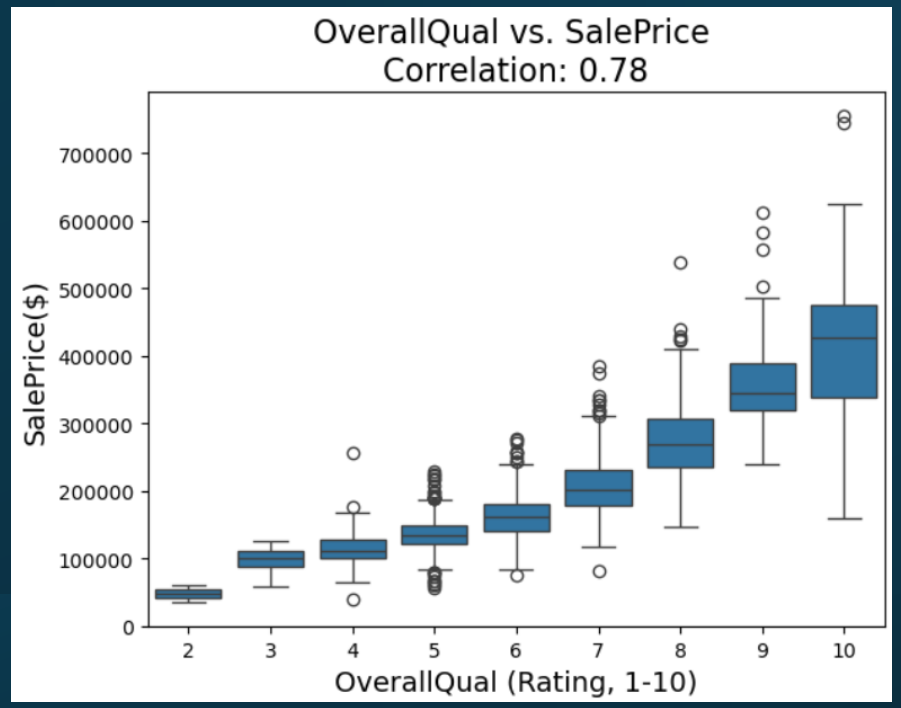


Features with the Strongest Correlation to SalePrice

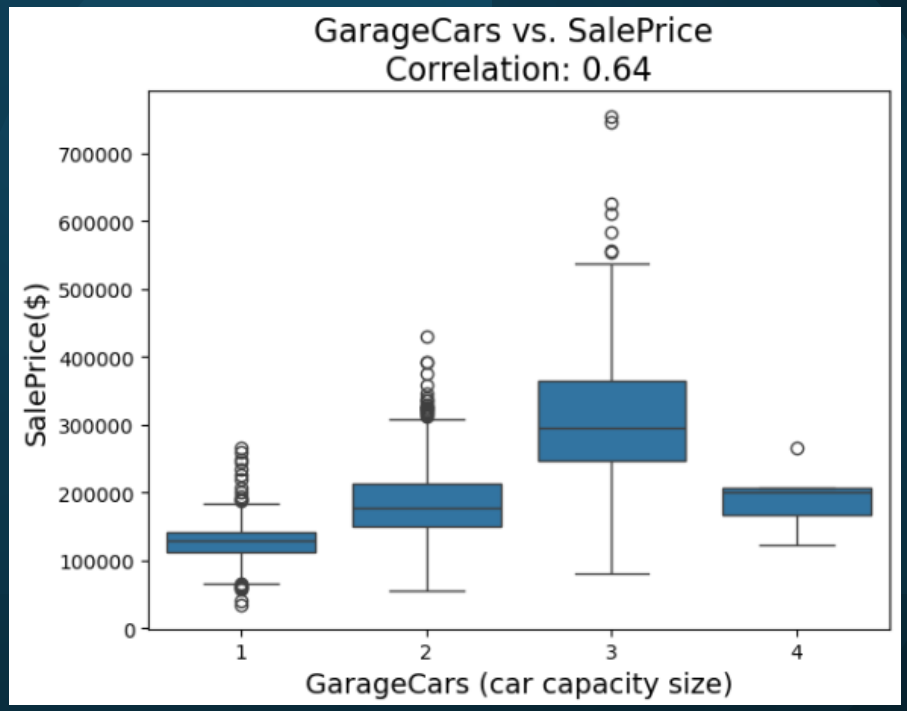


Data Exploration

OverallQual vs. SalePrice

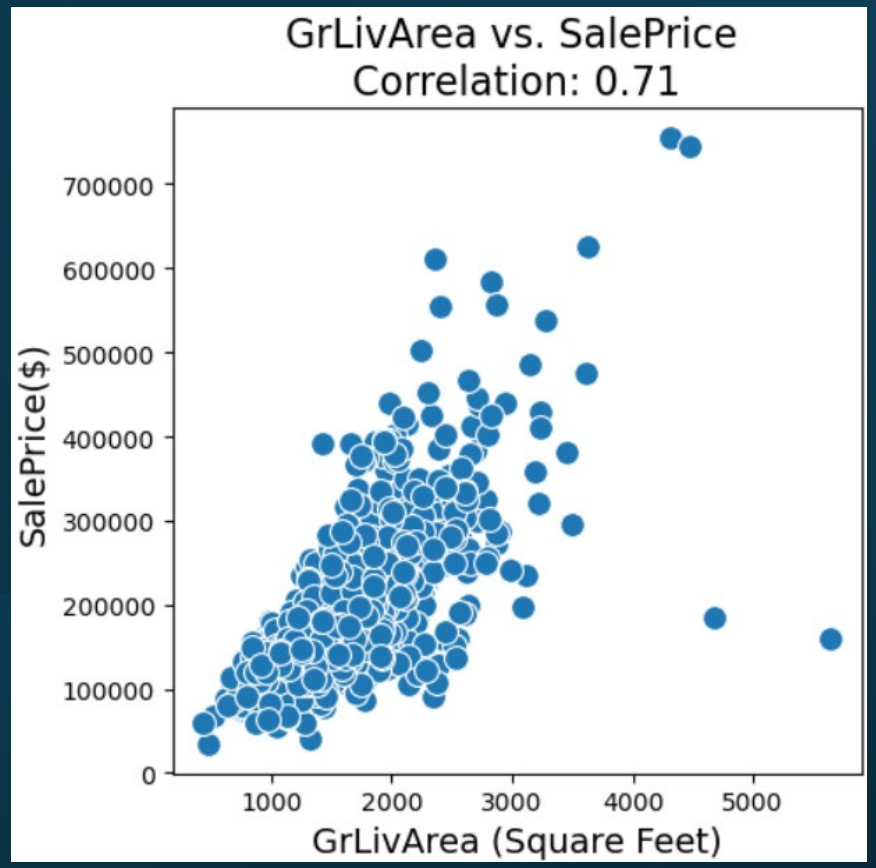


GarageCars vs. SalePrice

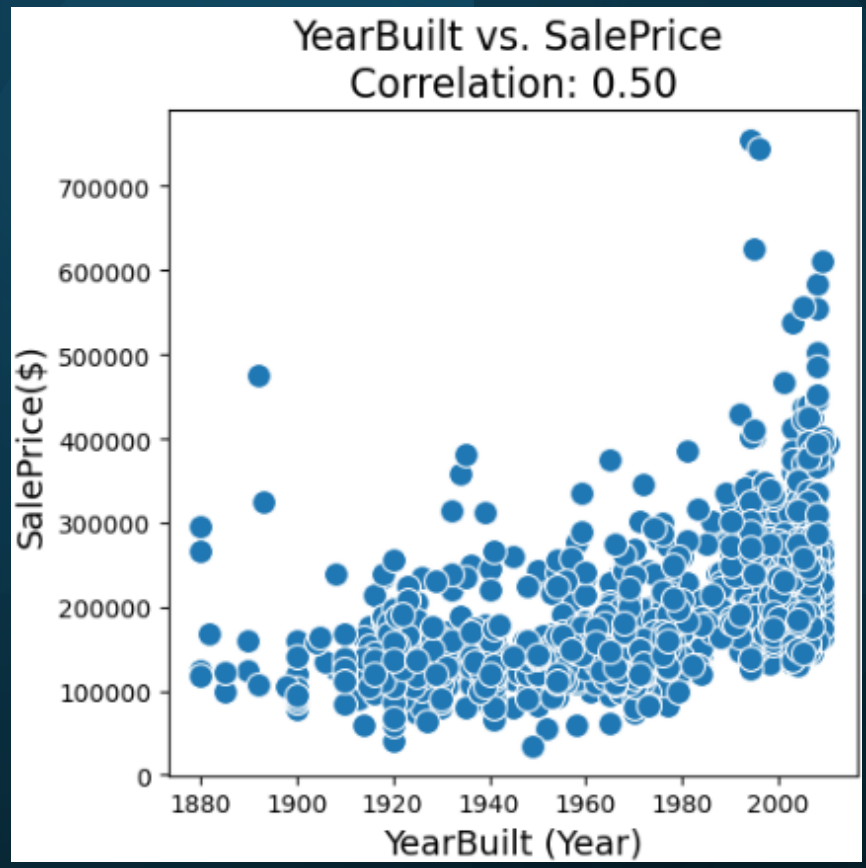


Data Exploration

GrLivArea vs. SalePrice



YearBuilt vs. SalePrice



Feature Engineering

Feature engineering

Scaler Selection Logic

Condition	Best Scaling Method
Year variables (YearBuilt, YrSold, YearRemodAdd)	Subtract Reference Year
Month variable (MoSold)	Sin-Cos Transformation
Ordinal variable	MinMaxScaler
Near-normal distributions with limited outliers	StandardScaler
Non-Gaussian but no extreme outliers	MinMaxScaler
Strong outliers, large spread, or moderate skew	RobustScaler
Highly skewed (e.g., >1.0), heavy-tailed distribution	Log Transformation

Feature engineering

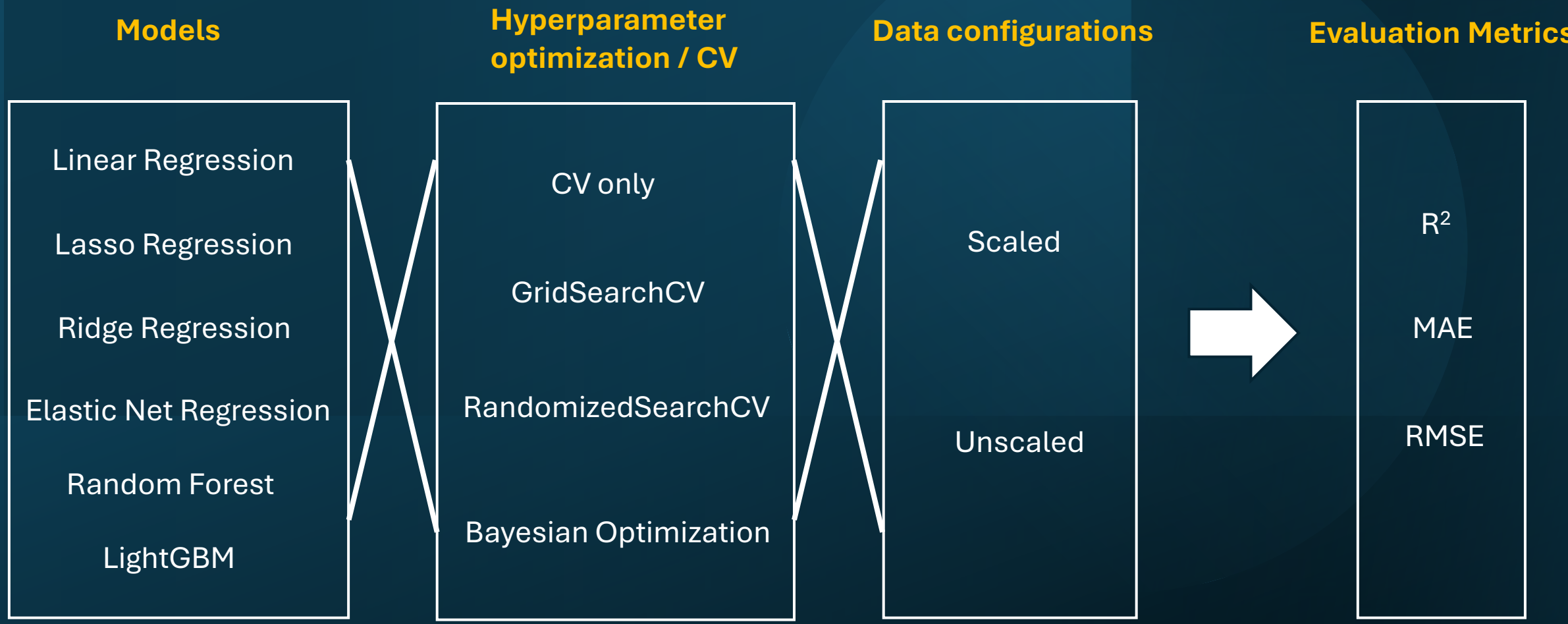
Scaler Assignment in Final

	MinMaxScaler	StandardScaler	RobustScaler	LogTransform	Reference_year	sin_cos_transform
0	BsmtFullBath		2ndFlrSF	WoodDeckSF	YearBuilt	MoSold
1	BsmtExposure		OverallCond	LotFrontage	YearRemodAdd	
2	HeatingQC		GarageYrBlt	LotArea	YrSold	
3	GarageFinish		PavedDrive	1stFlrSF		
4	OverallQual		BsmtCond	EnclosedPorch		
5	KitchenQual		ExterCond	GrLivArea		
6	FullBath		TotRmsAbvGrd	TotalBsmtSF		
7	GarageCars		BsmtUnfSF	SalePrice		
8	BsmtQual		GarageArea	BsmtFinSF2		
9	HalfBath		BedroomAbvGr	OpenPorchSF		
10	ExterQual		GarageQual	MasVnrArea		
11			Fireplaces	BsmtFinSF1		

Machine Learning Modeling

Machine Learning Modeling

Machine Learning Modeling

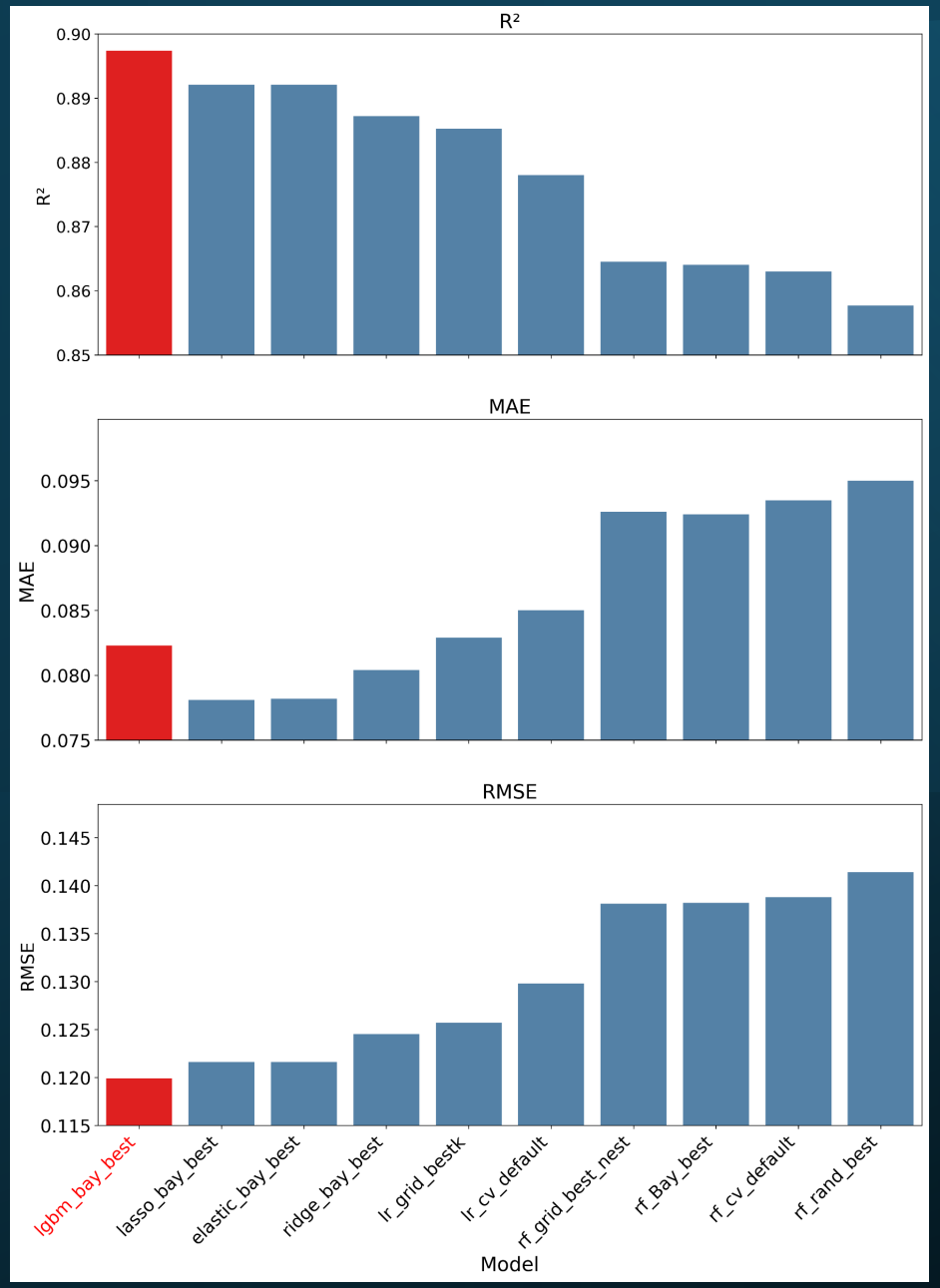


Machine Learning Modeling

Model Comparison

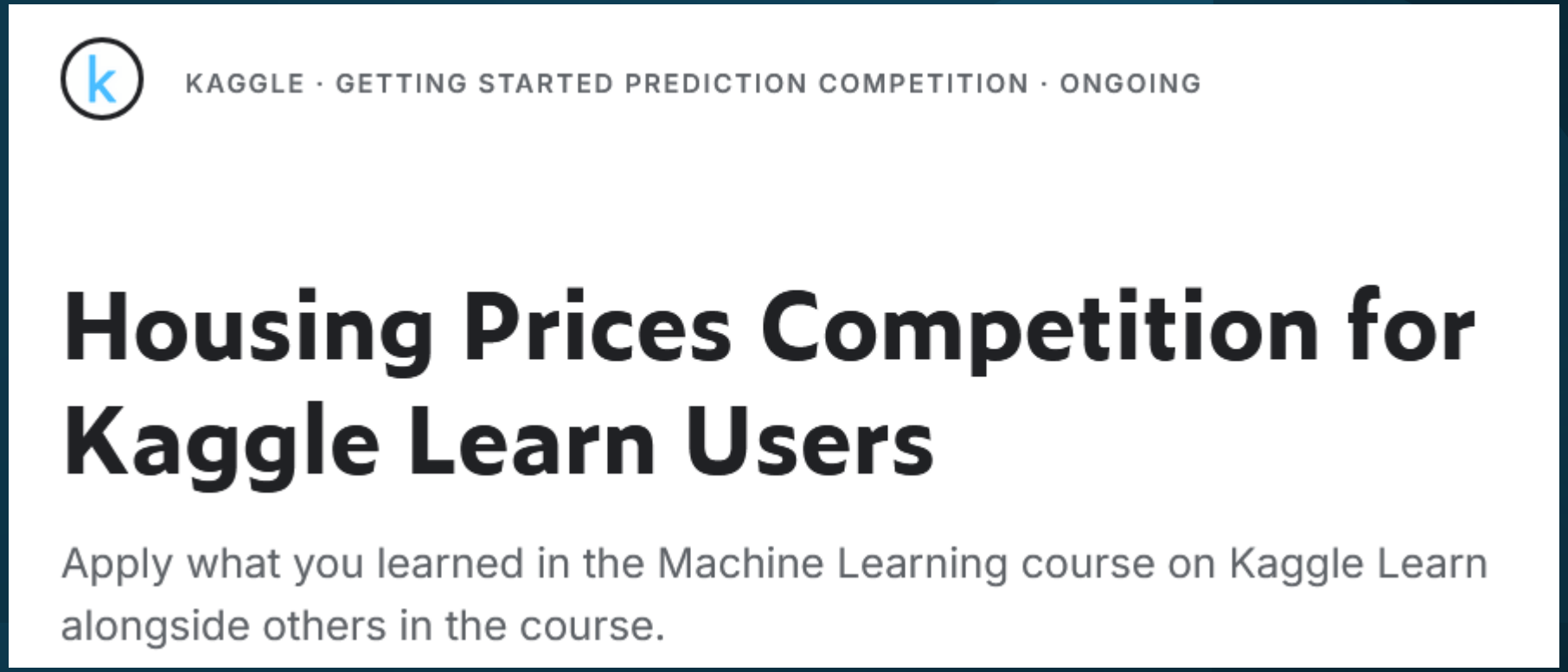
	Model	Hyperparameter Opt. / CV	Training Dataset	Model ID	R^2	MAE	RMSE
1	LightGBM	Bayesian Optimization	Scaled	lgbm_bay_best	0.8974	0.0823	0.1199
2	Lasso regression	Bayesian Optimization	Scaled	lasso_bay_best	0.8921	0.0781	0.1216
3	Elastic Net Regression	Bayesian Optimization	Scaled	elastic_bay_best	0.8921	0.0782	0.1216
4	Ridge regression	Bayesian Optimization	Scaled	ridge_bay_best	0.8872	0.0804	0.1245
5	Linear regression	GridSearchCV	Scaled	lr_grid_bestk	0.8852	0.0829	0.1257
6	LightGBM	Bayesian Optimization	Unscaled	lgbm_unscale_bay_best	0.8793	15844.1	27241.4
7	Linear regression	CV	Scaled	lr_cv_default	0.878	0.085	0.1298
8	Random Forest	GridSearchCV	Scaled	rf_grid_best_nest	0.8645	0.0926	0.1381
9	Random Forest	Bayesian Optimization	Scaled	rf_Bay_best	0.864	0.0924	0.1382
10	Random Forest	CV	Scaled	rf_cv_default	0.863	0.0935	0.1388
11	Random Forest	RandomizedSearchCV	Scaled	rf_rand_best	0.8577	0.095	0.1414
12	Random Forest	Bayesian Optimization	Scaled	rf_unscale_Bay_best	0.8535	17502.9	30053.8

Machine Learning Modeling



The model **lgbm_bay_best** was selected as the best performer and subsequently used for the downstream analysis.

Prediction on Testing data



Ranked 392nd out of 24,835 submissions, placing in the top 1.5%

More Ideas to Improve the Model in Future:

1. Explore top-performing public notebooks for modeling strategies and feature engineering ideas;

2. Experiment with other advanced ensembling techniques (e.g., combining XGBoost, and CatBoost);

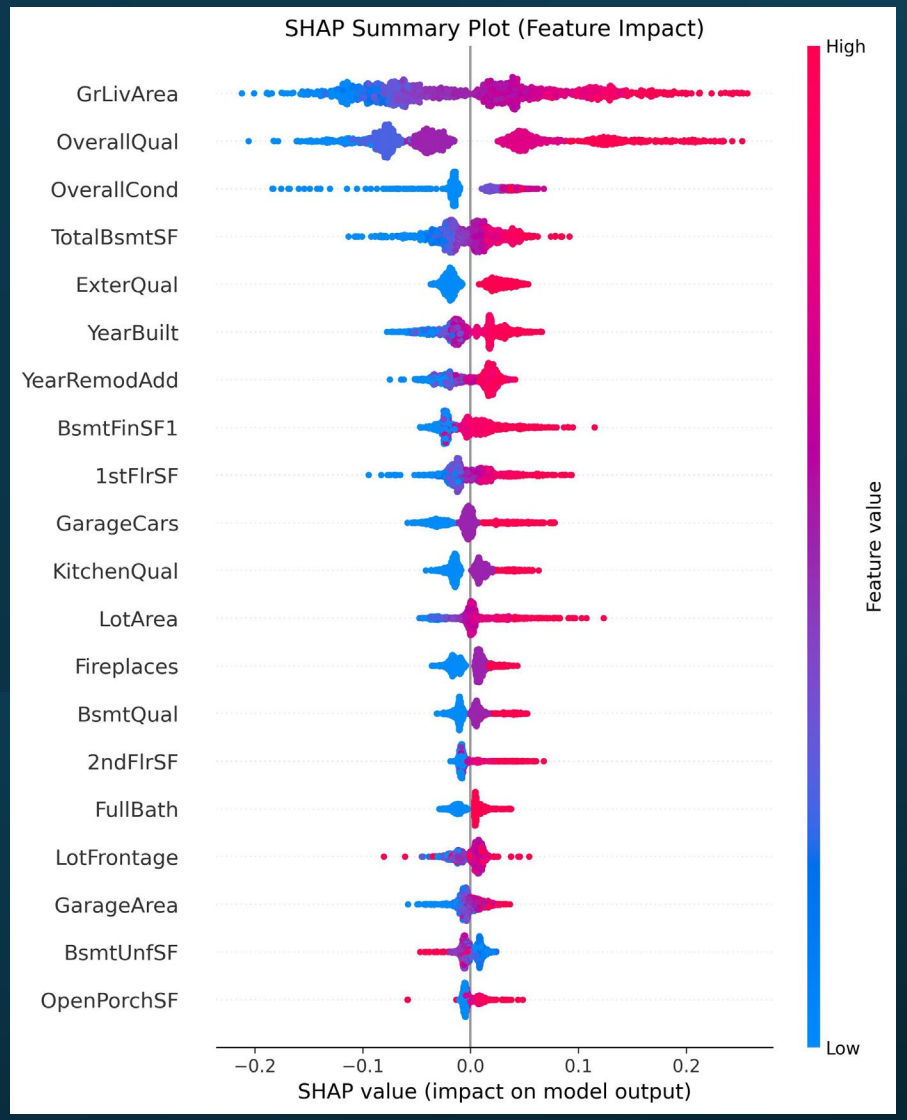
3. Implement cross-validation with out-of-fold predictions to better control overfitting;

4. Conduct deeper feature engineering, which is often a key factor distinguishing top-tier solutions.

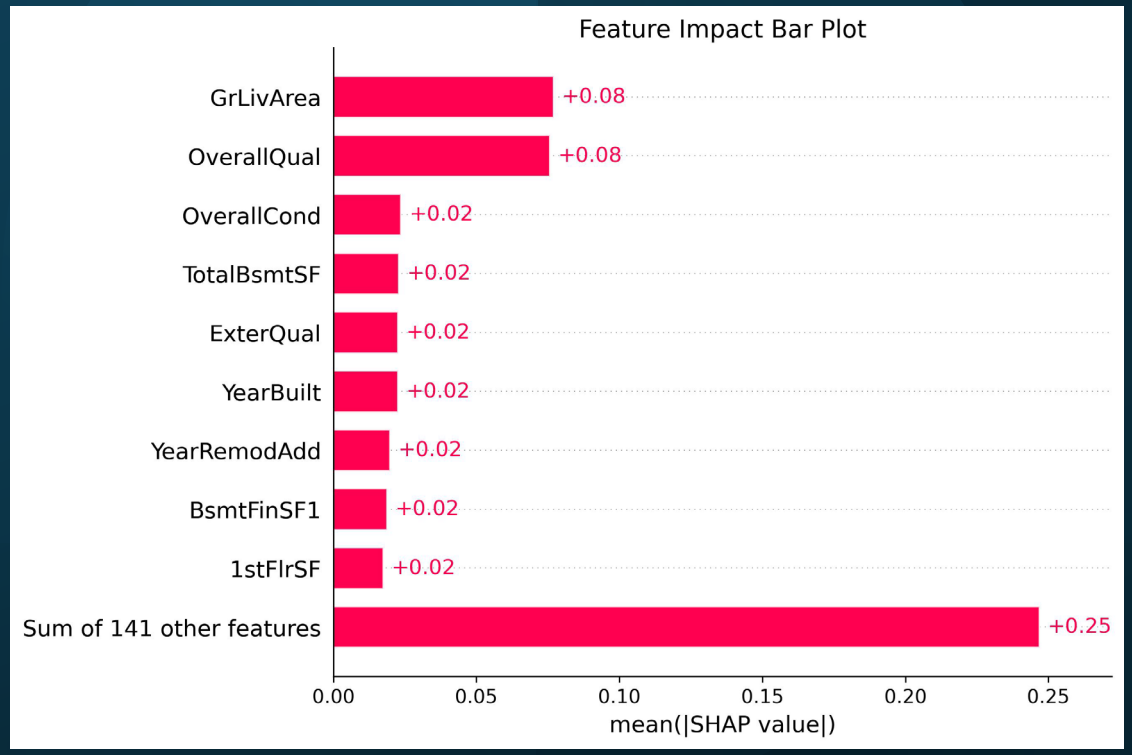
Feature Importance

Feature Importance

SHAP Summary Plot

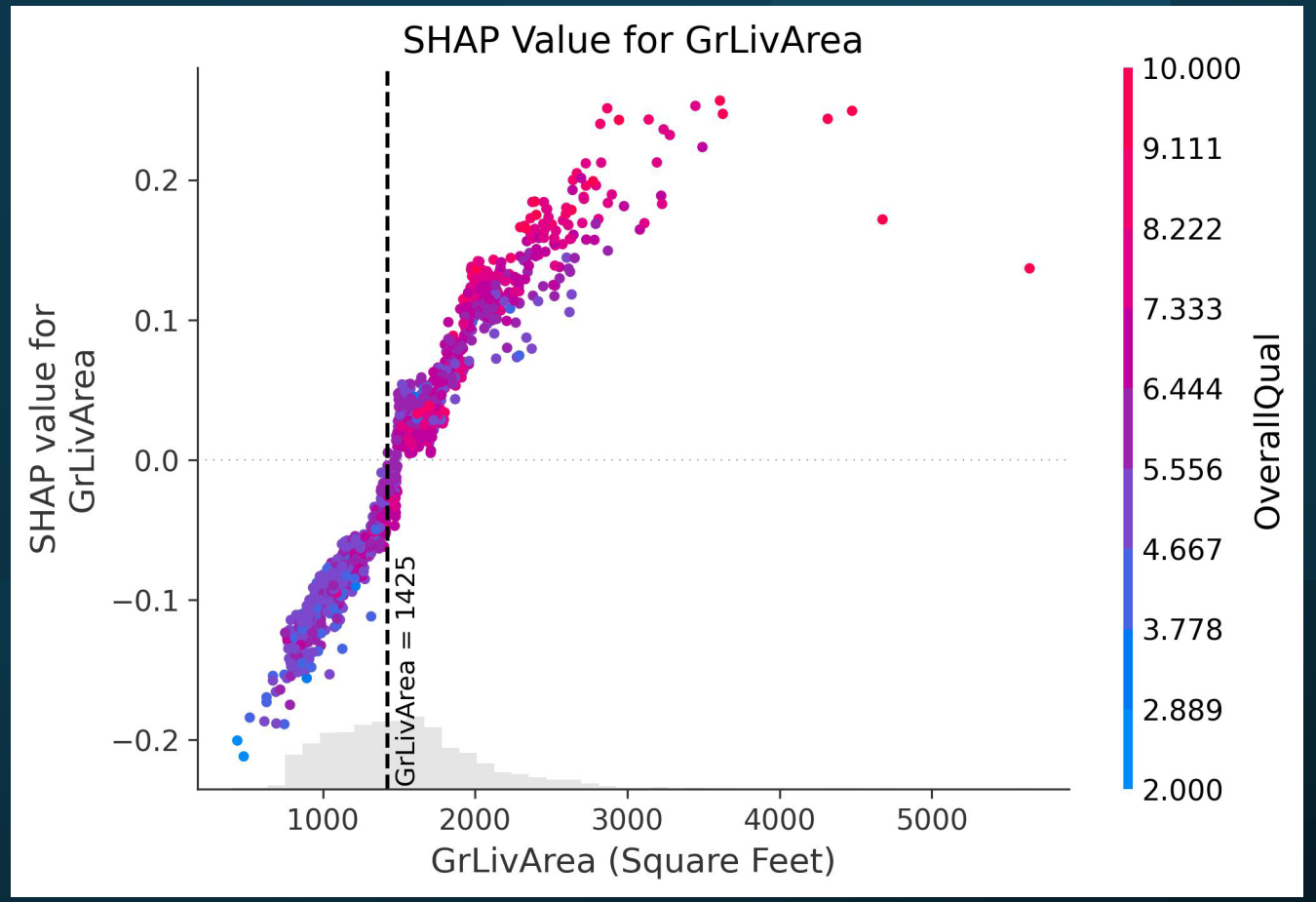


Feature Impact Bar Plot



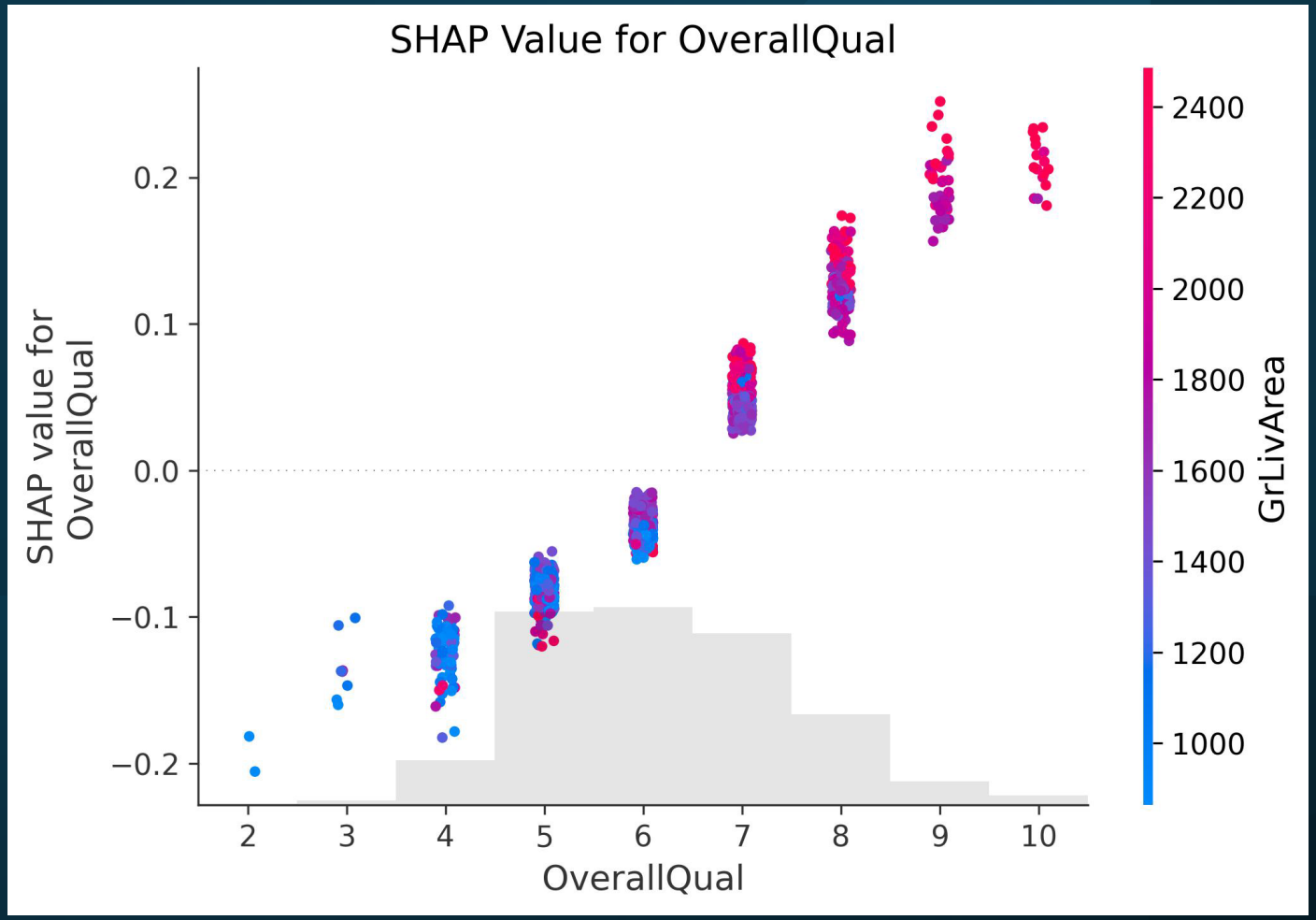
SHAP Dependence Plots for Key Features

GrLivArea



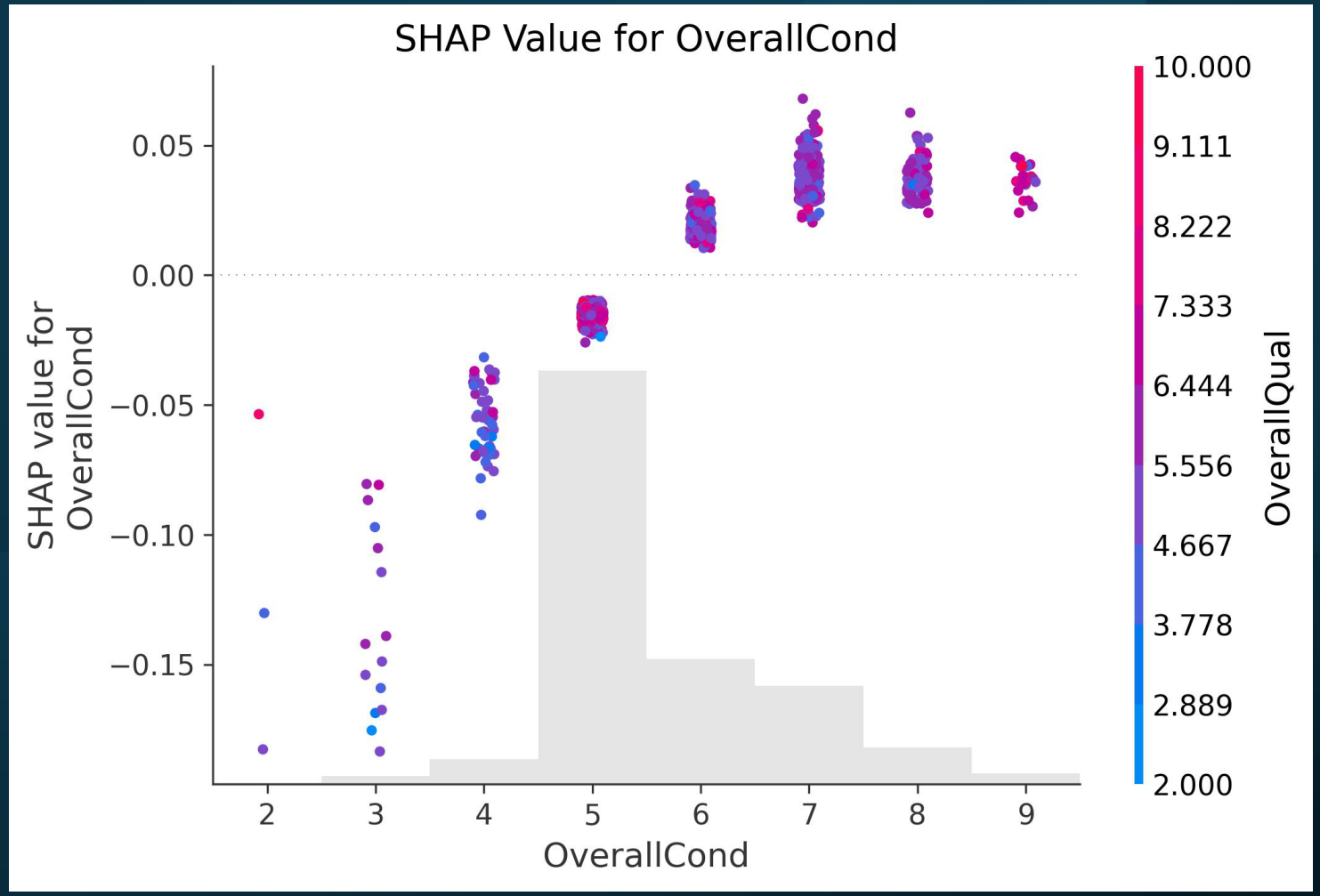
SHAP Dependence Plots for Key Features

OverallQual



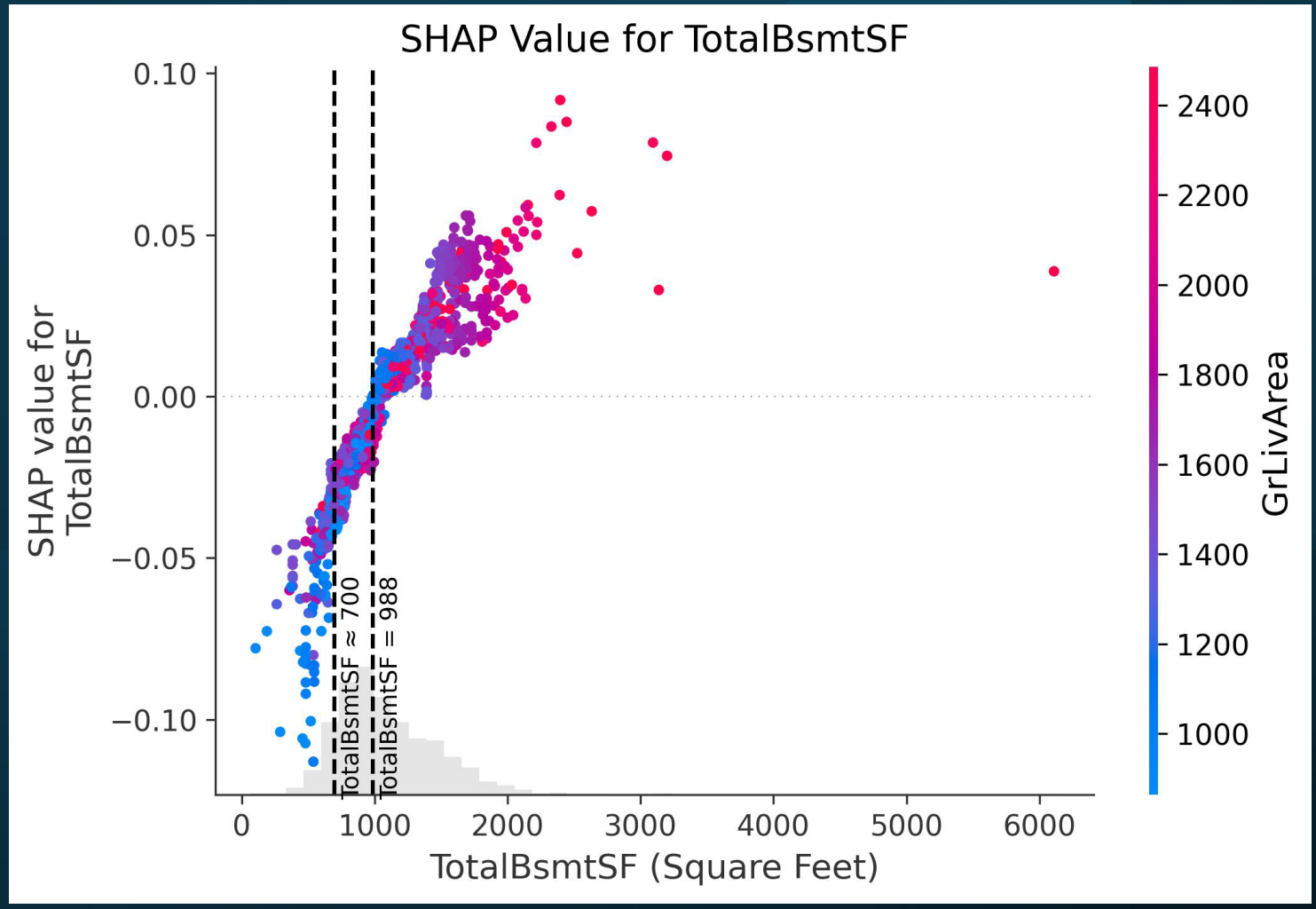
SHAP Dependence Plots for Key Features

OverallQual



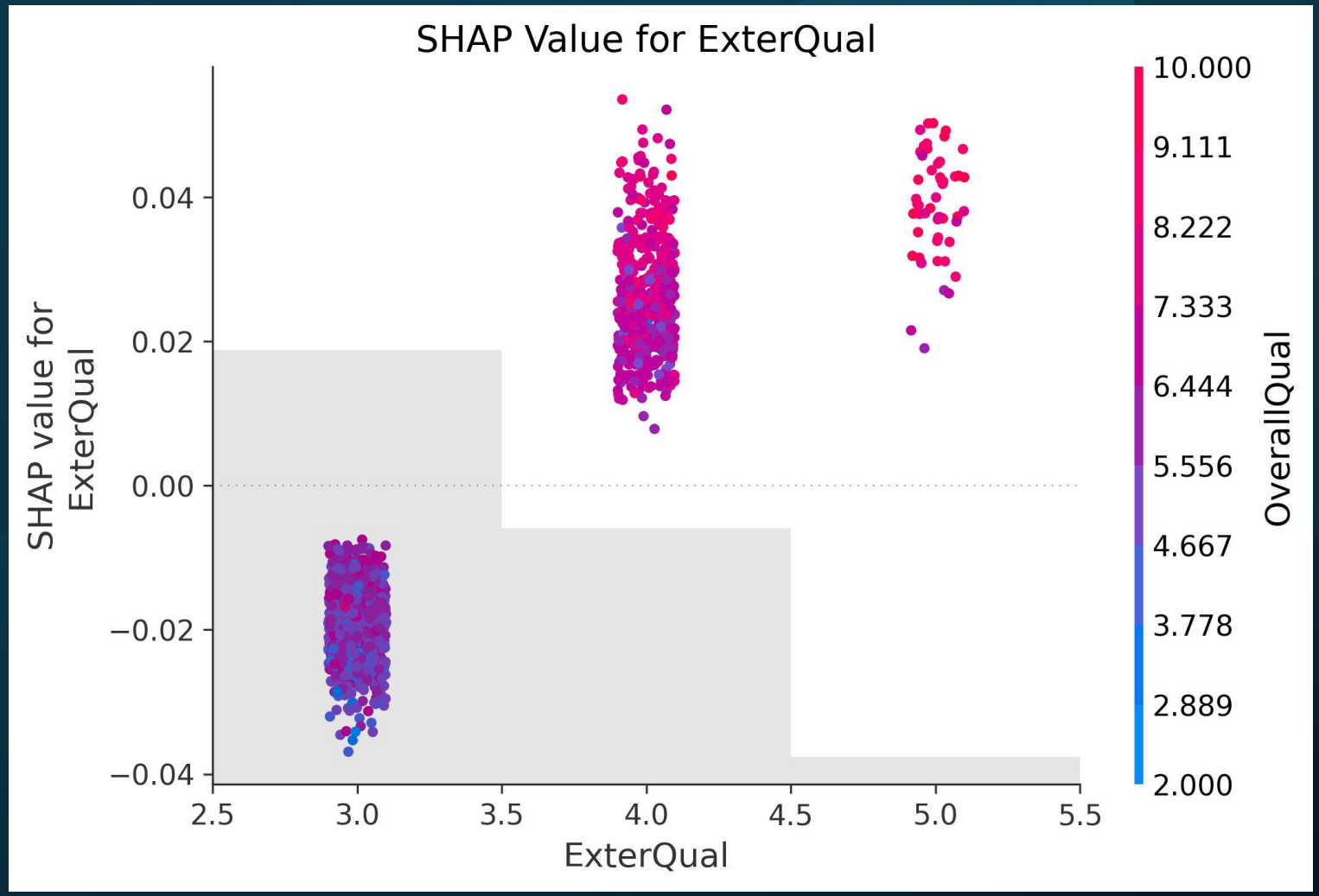
SHAP Dependence Plots for Key Features

TotalBsmtSF



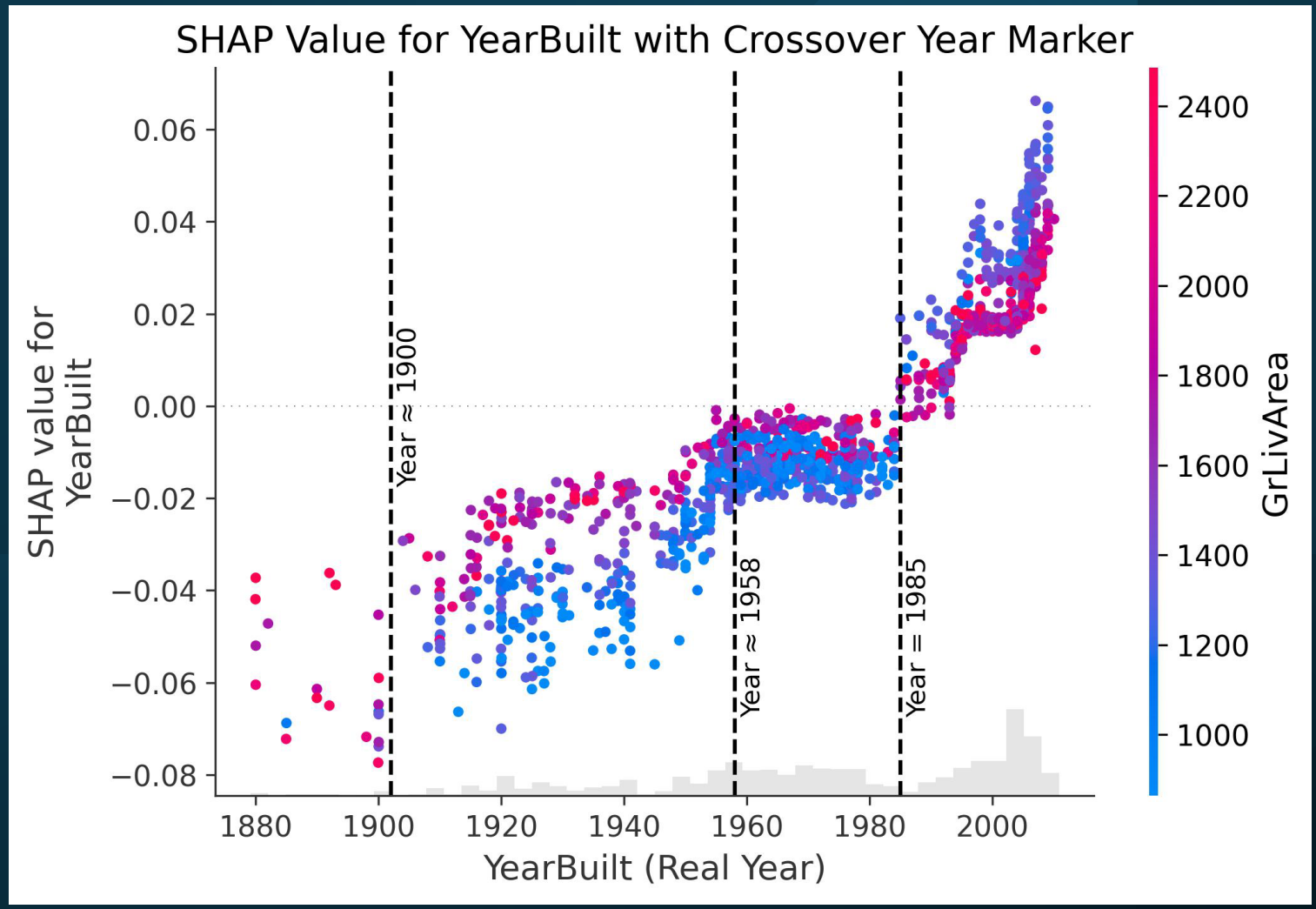
SHAP Dependence Plots for Key Features

ExterQual



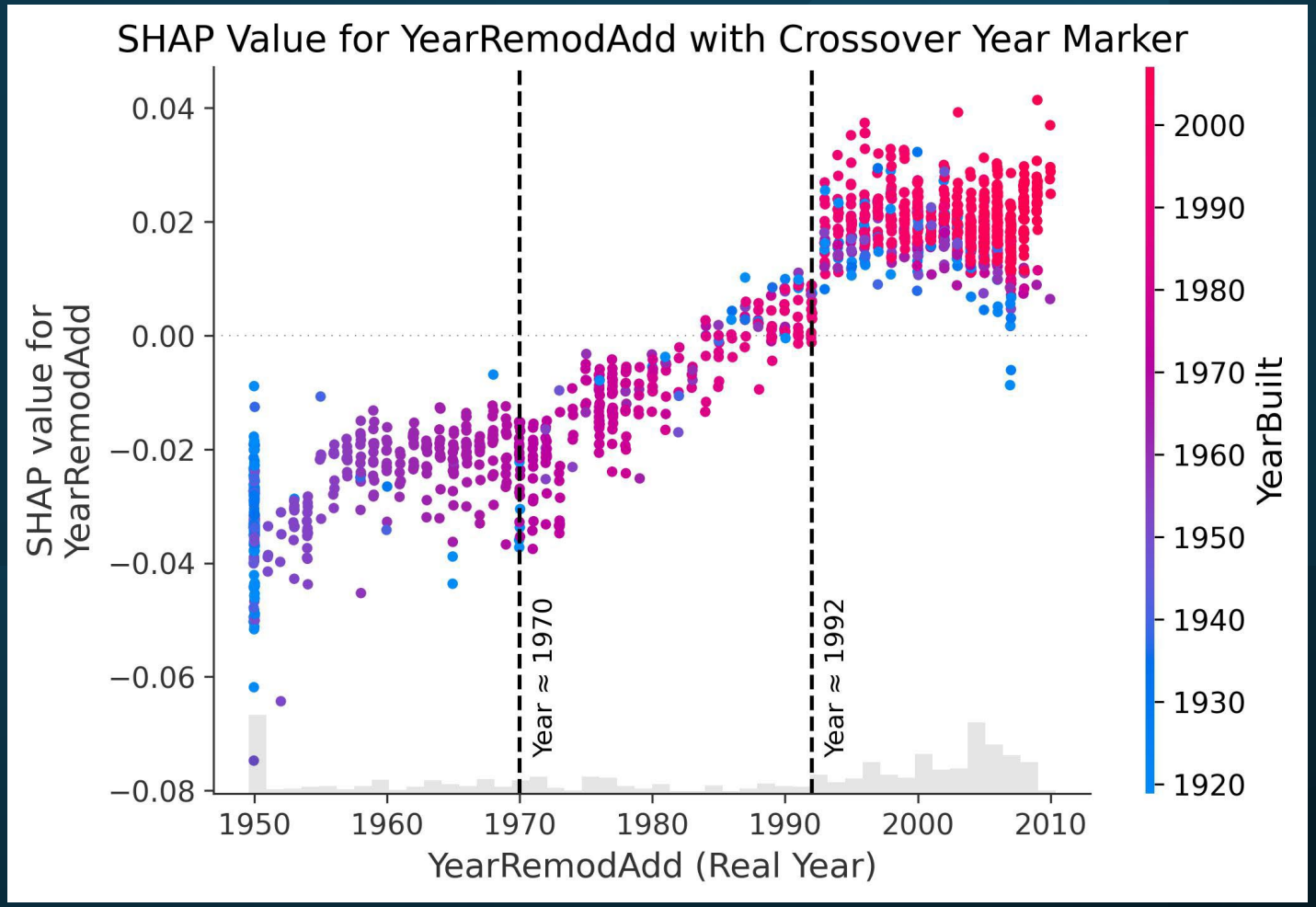
SHAP Dependence Plots for Key Features

YearBuilt



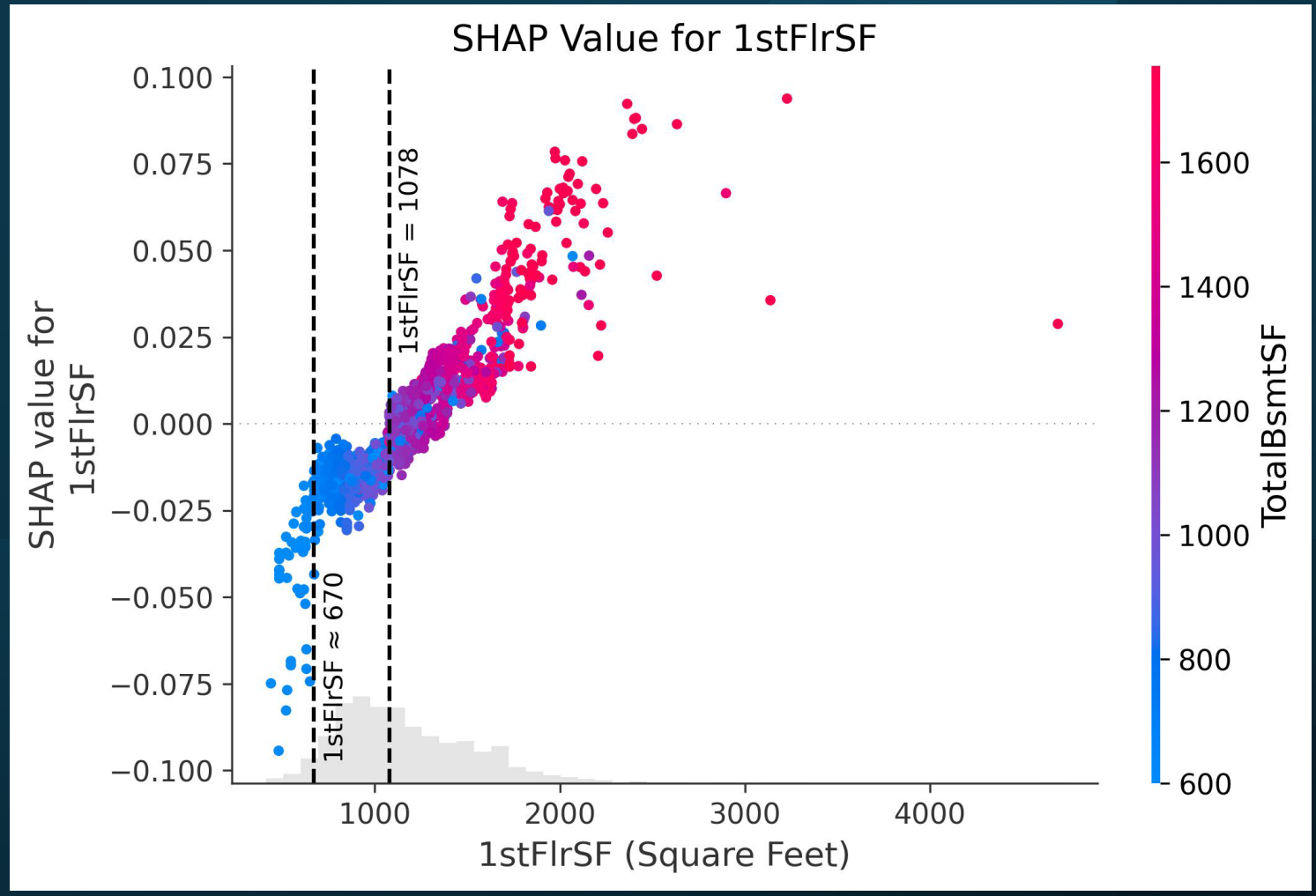
SHAP Dependence Plots for Key Features

YearRemodAdd



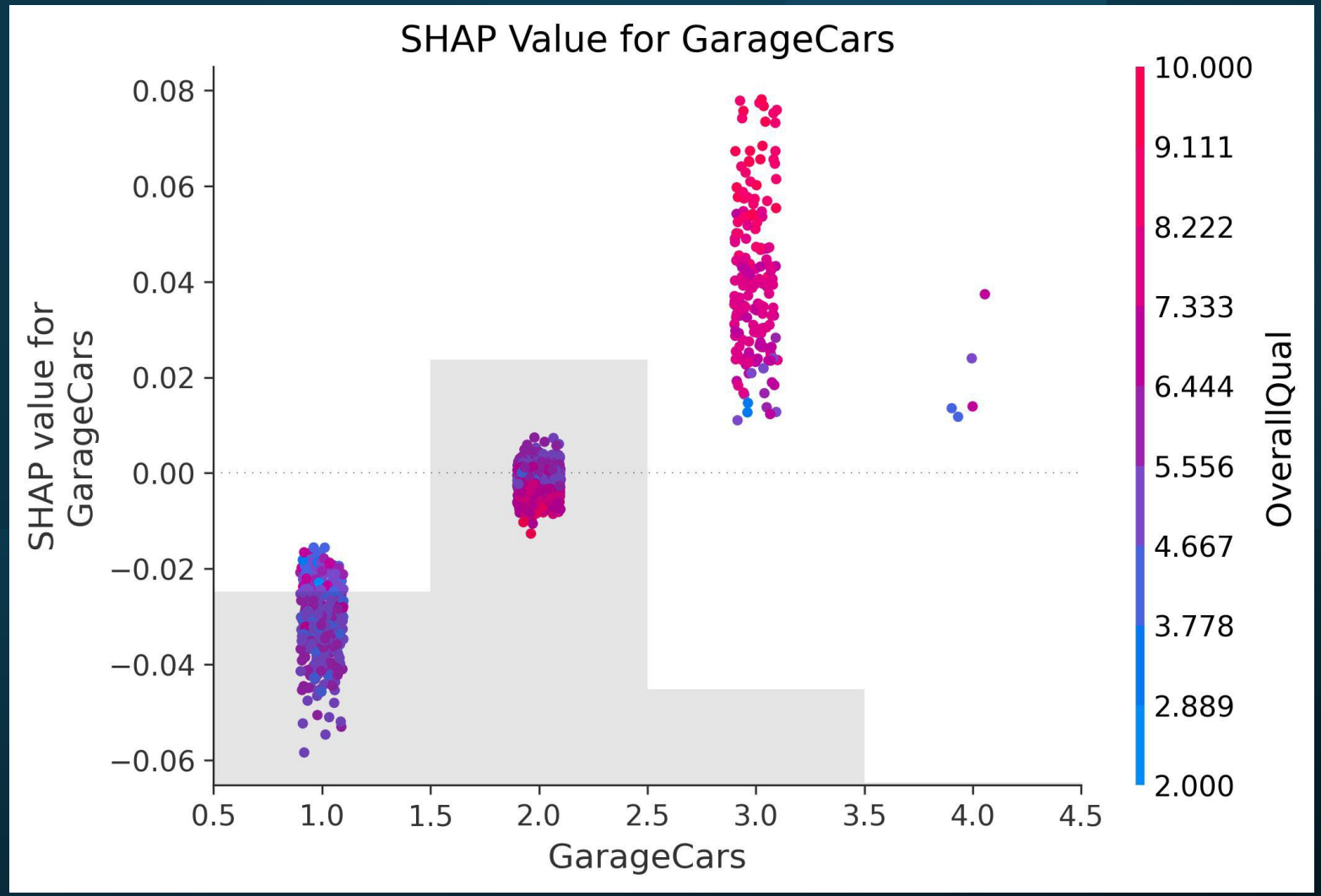
SHAP Dependence Plots for Key Features

1stFlrSF



SHAP Dependence Plots for Key Features

GarageCars



Conclusions

1. Out of 12 models created, LightGBM with Bayesian optimization provides the best results.
2. Insights from SHAP value analysis:
    a. Living Space Is the Most Powerful Predictor
    b. Quality Matters More Than Condition
    c. Garage Capacity Shows Value Saturation
    d. First Floor Area Shows Nonlinear Influence
    e. YearBuilt – Age of Home Matters, Especially for Small Homes
    f. YearRemodAdd – Remodeling Improves Value, But Only Up to a Point
    g. Interacting Effects Are Crucial

Acknowledgement

Benjamin Bell

Thank you