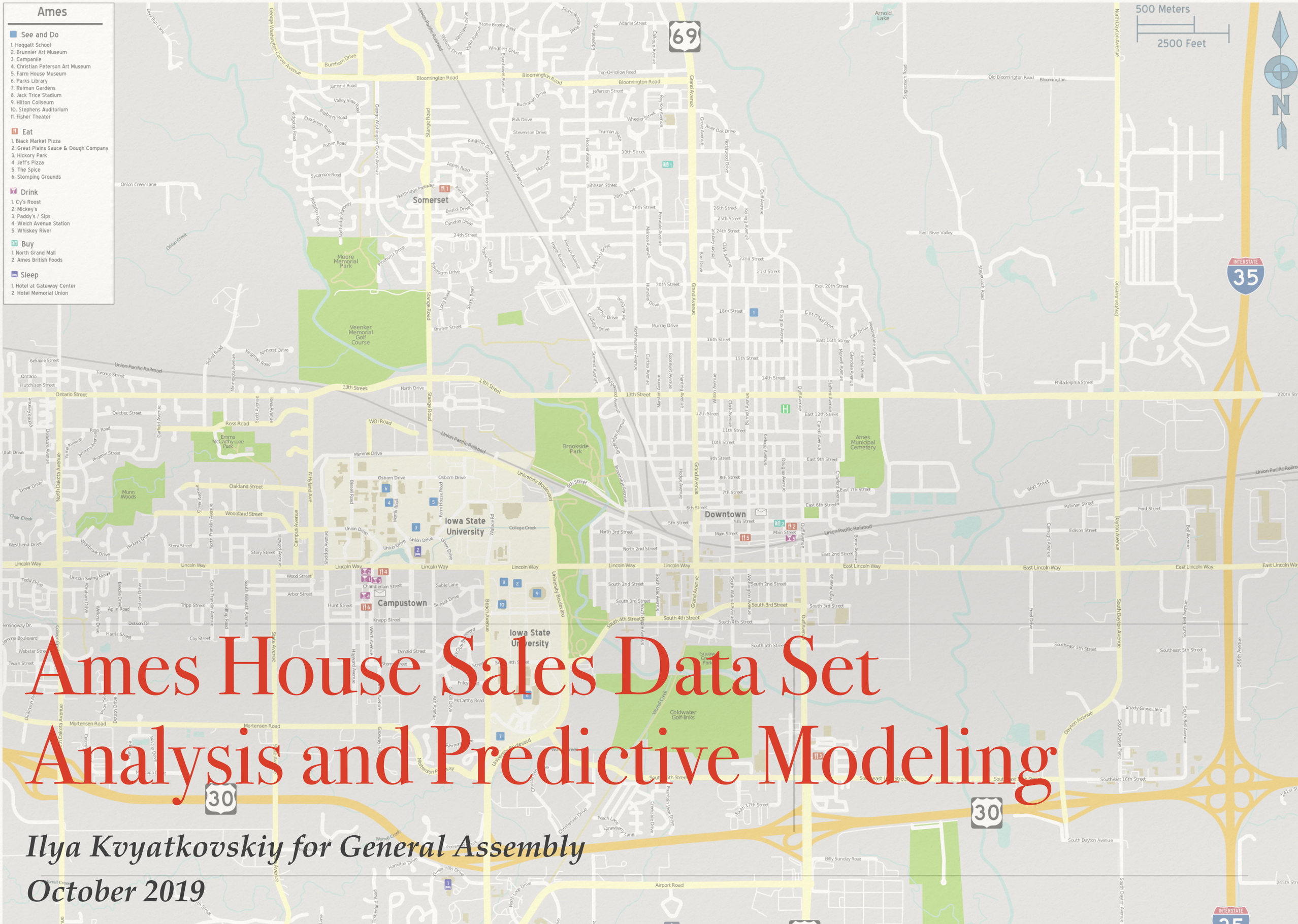


Ames House Sales Data Set Analysis and Predictive Modeling

Ilya Kvyatkovskiy for General Assembly

October 2019

Problem Statement

Demonstrate use of different linear regression techniques analyzing Ames, IA housing data set.

With necessary assumptions being made, analyze various linear regression methods performance, identify the best performing model and evaluate its performance.

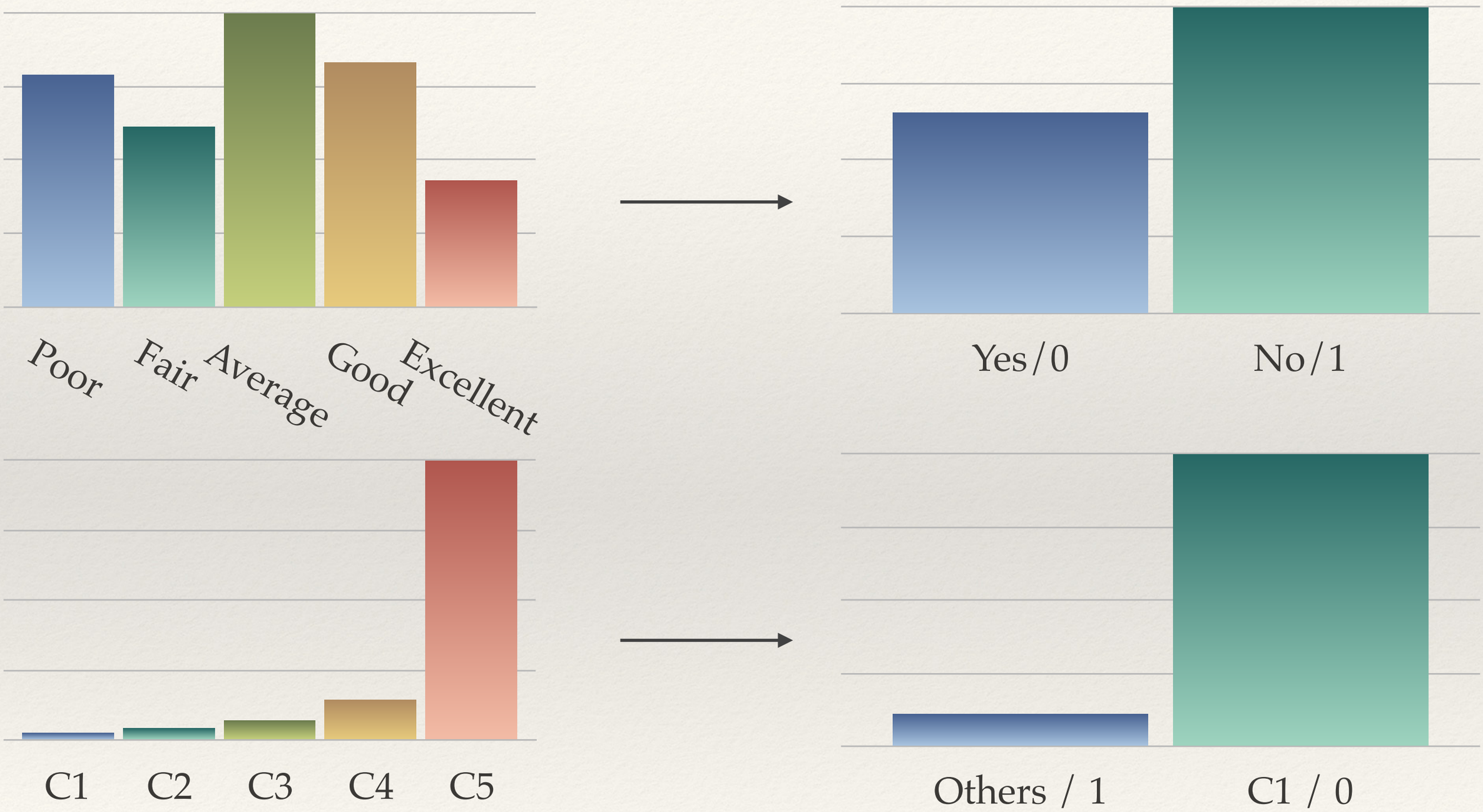
While working on project participate in group challenge at ***Kaggle*** *testing the model performance on a testing data sets with unknown target values.*

Project Workflow

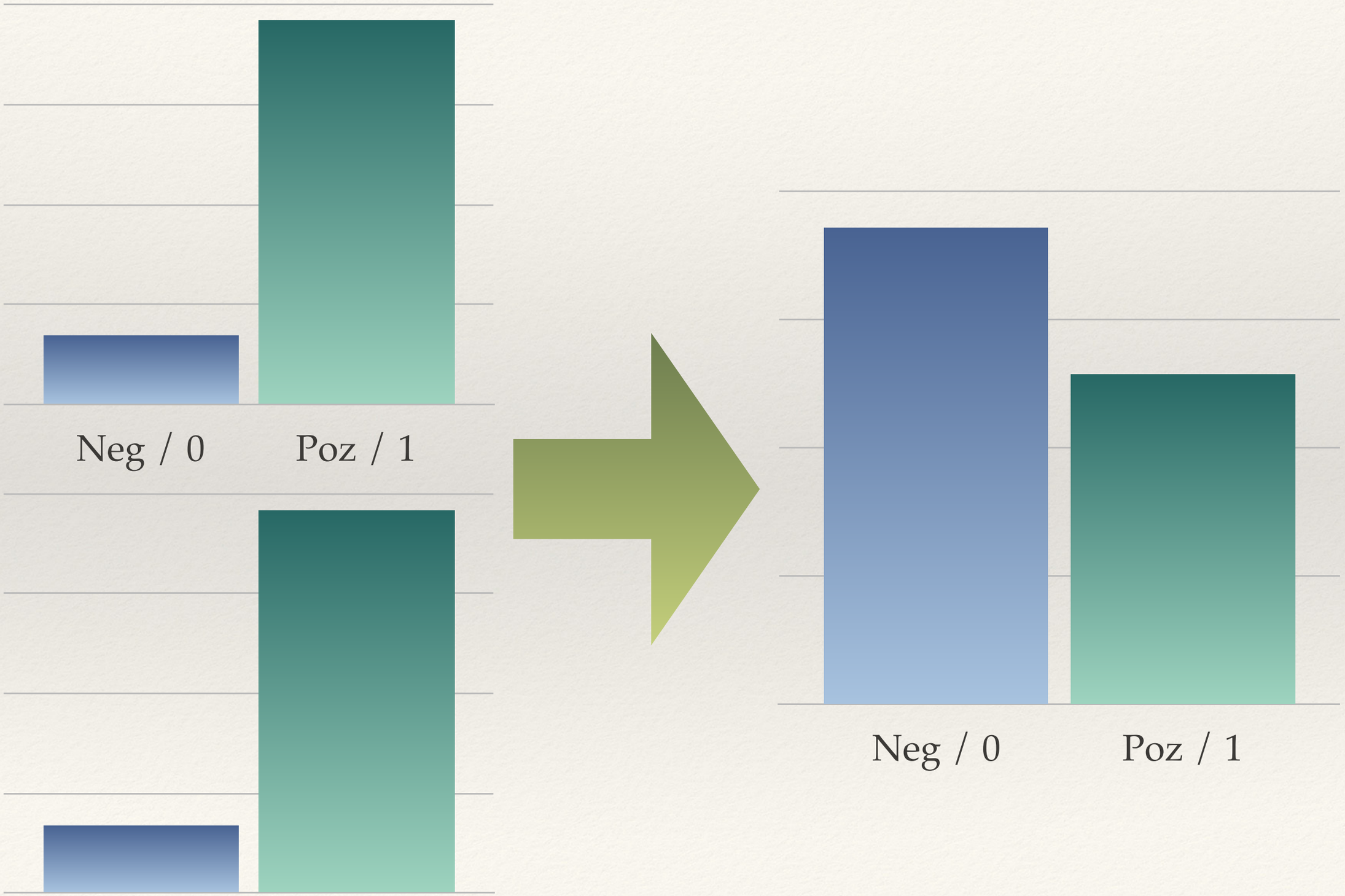
1. Initial Data Analysis
2. Data Cleaning
3. Feature Engineering
4. Data Pre-processing
5. Modeling
    - Simple Multilinear Regression
    - Power Transformation
    - Polynomial Features
    - Ridge Regression
    - Lasso Regression
    - ElasticNet Regression
6. Conclusions and Recommendations

Data Cleaning

Feature Engineering

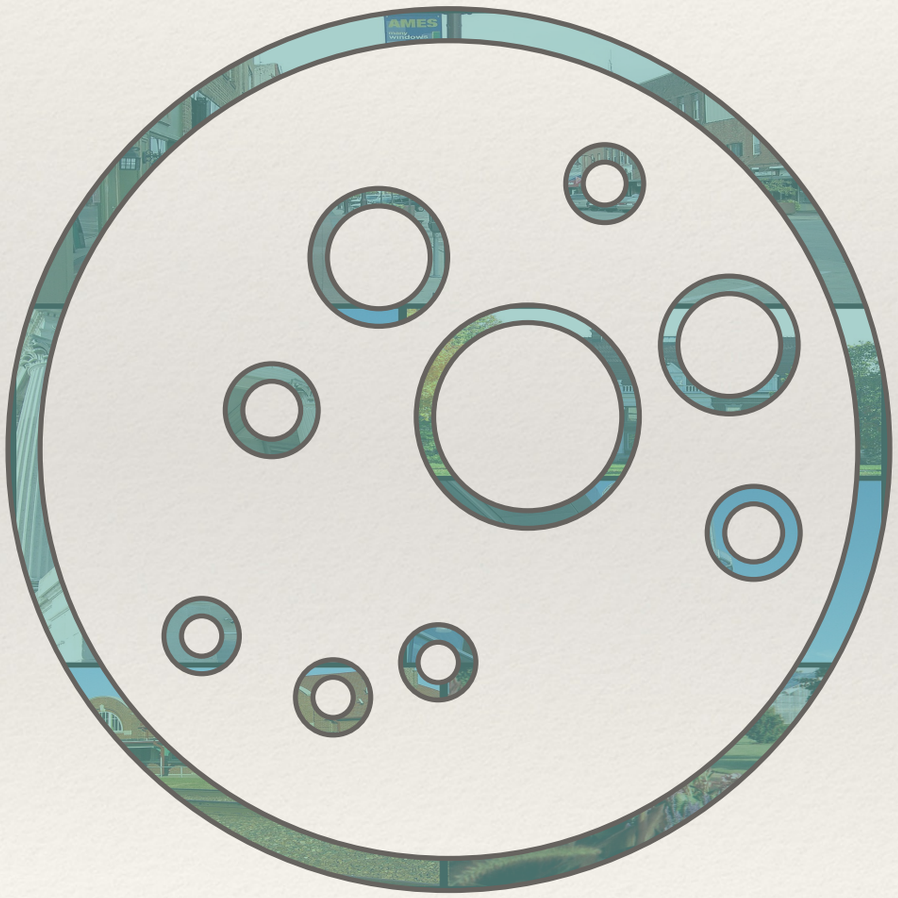


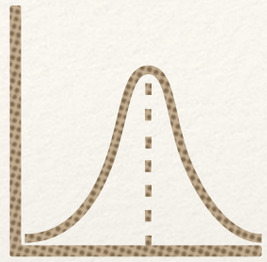
Feature Engineering

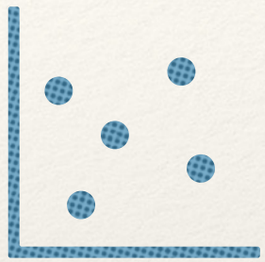


Feature Matrices

Potential Predictors (29)

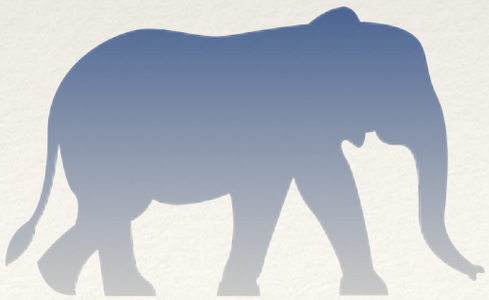


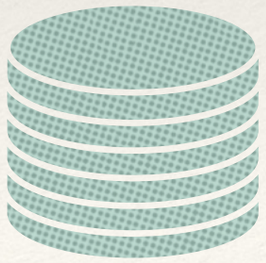




Features Set (12)
- Filtered
- Minimized collinearity

Extended Features Set (29)
- Not filtered
- Strong collinearity



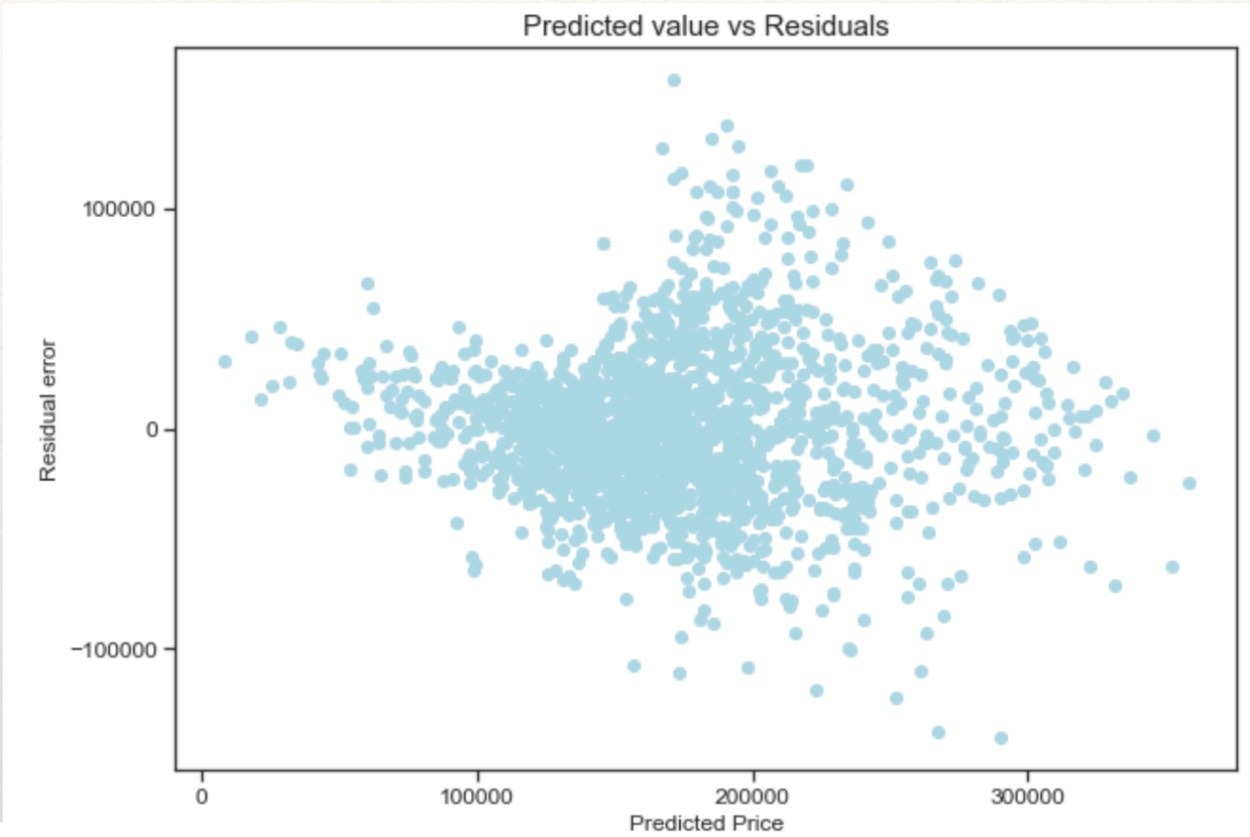


Simple MLR

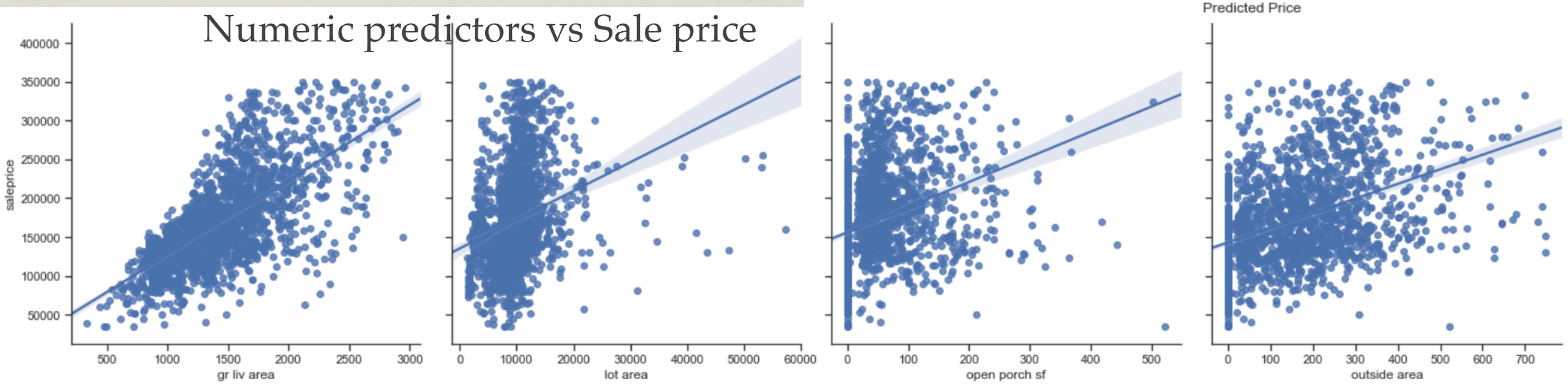
Cross-validated RMSE = 0.6743

Extended Features Set

Cross-validated RMSE = **0.8762**



Numeric predictors vs Sale price



Power Transformations & PolynomialFeatures

Polynomial Features (Power = 2)	
Features Set	Cross-validated RMSE = 0.7128
Extended Features Set	Cross-validated RMSE = 0.8068

Power Transformations			
Square root out of all numeric predictors RMSE = 0.7123	Square root out of all numeric predictors, and		
	[Lot Area]exp(1/3)	[Lot Area]exp(1/4)	[Lot Area]exp(1/10)
	0.6912	0.6929	0.6932

Ridge, Lasso and ElasticNet

Ridge Regression (Extended Features)		
Default	Alpha = 1.0	Cross-validated RMSE =0.88622
Optimized	Alpha = 2.15	Cross-validated RMSE = 0.88620
Lasso Regression		
Default	Alpha = 1.0	Cross-validated RMSE = **0.88623**
Optimized	Alpha = 1.0	Cross-validated RMSE = **0.88623**
ElasticNet Regression		
Optimized	Alpha = 0.5	Cross-validated RMSE = 0.87689

Best Performing Model Evaluation

Lasso Regression Model with Cross-Evaluated RMSE score of 0.88623

Evaluation method: Train-Test-Split	Training Set RMSE	0.8919
	Testing Set RMSE	0.8633

Conclusion: the two RMSE scores for training and testing sets are quite similar, which indicated a quite well-balanced model. This model could be expected to generalize quite well on unknown data.

Conclusions

- Model with the highest R2 score of 0.88623 is a Lasso regression model fitted with a multitude of original predictors each having an absolute correlation with the target value of no less than 0.2

- When evaluating this model using train-test-split procedure I received a R2 score of 0.8919 for the training set and a R2 score of 0.8633 for a testing set, which indicates a well-balanced model and gives a hope the model will work well on unknown data

- Still, Kaggle predictions on a set of unknown data have a quite high RMSE which most probably results from two things:
    - Initial assumptions while possible predictors analysis/choice
    - Quite high level of multicollinearity between chosen predictors