Where does the value of your home come from? Using SHAP to see nuances in home sale price predictions in Ames, Iowa

Natalie Stier

NYCDSA

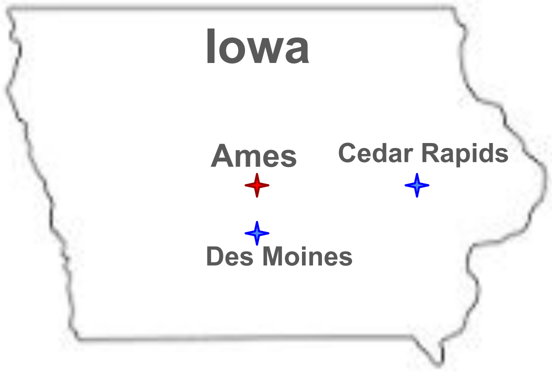
Machine Learning Project
August 9, 2024

1

Introduction: Ames Housing Data

The Ames, Iowa housing data was assembled in 2011 by Dean De Cock.

It is as an alternative to the 1978 Boston Housing Data Set which he had worked with as a master's student at Iowa State University, located in Ames.

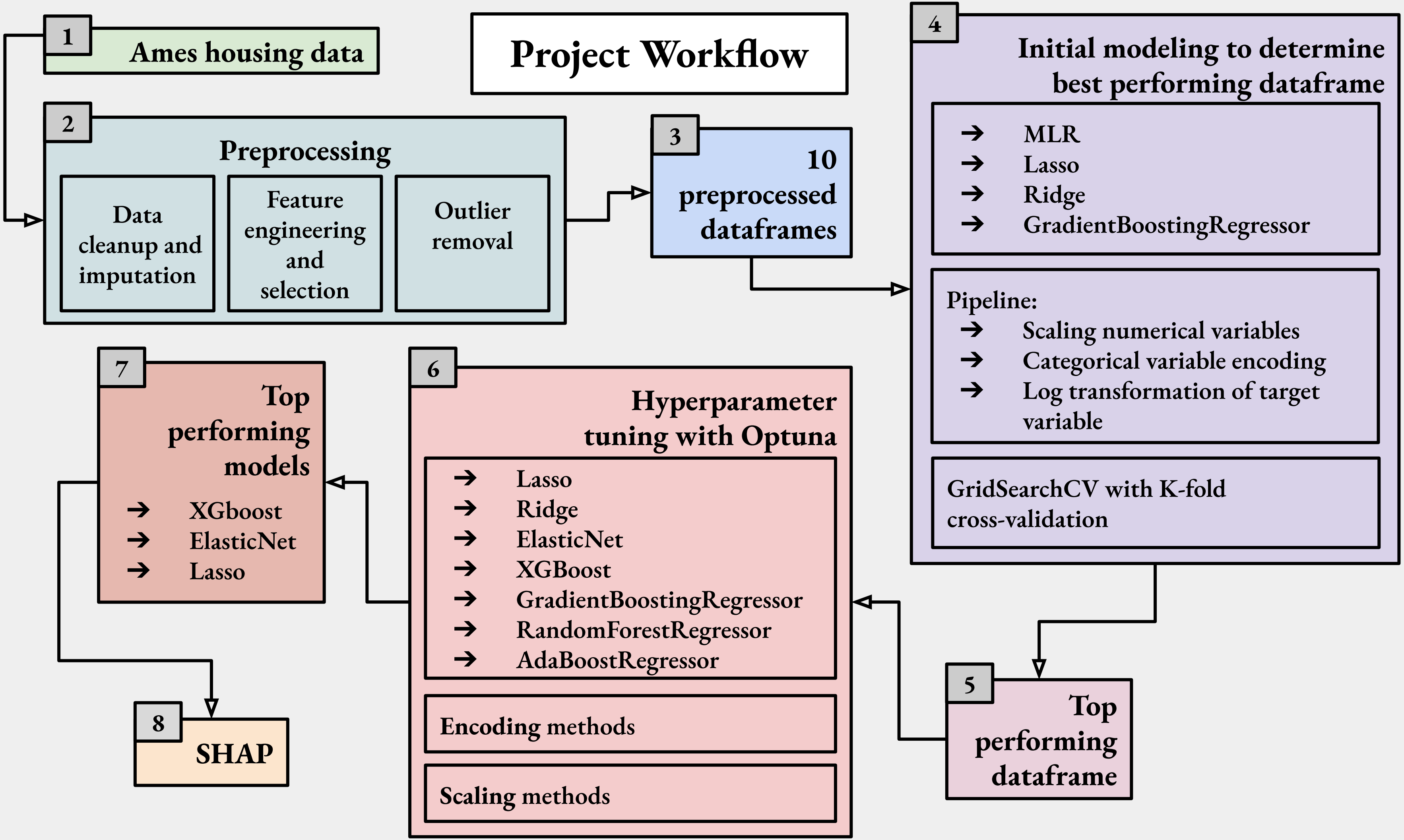


The data set for this project includes 2580 observation and 81 columns

➔ 79 features (nominal, ordinal, continuous, and discrete variables)

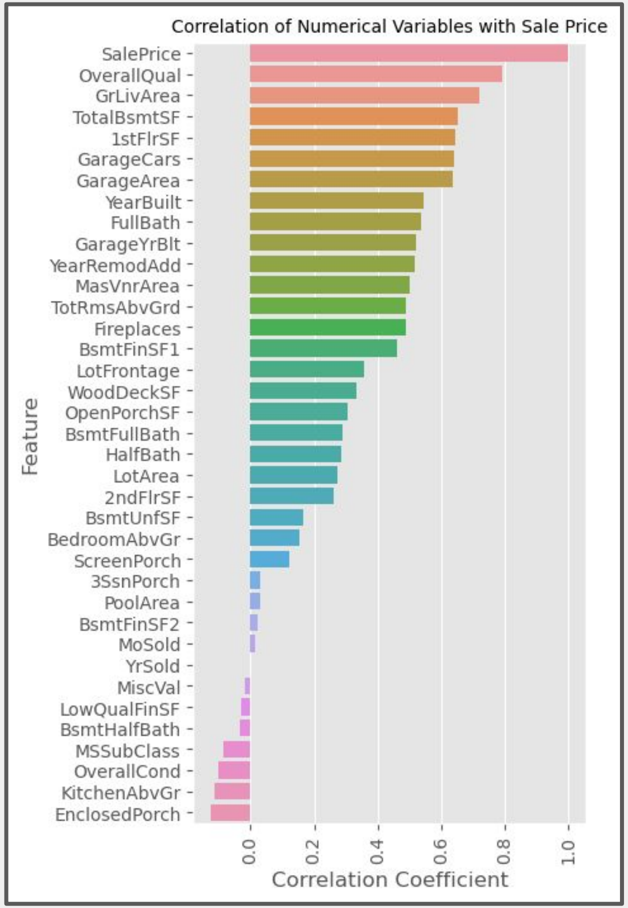
➔ SalePrice, the target variable

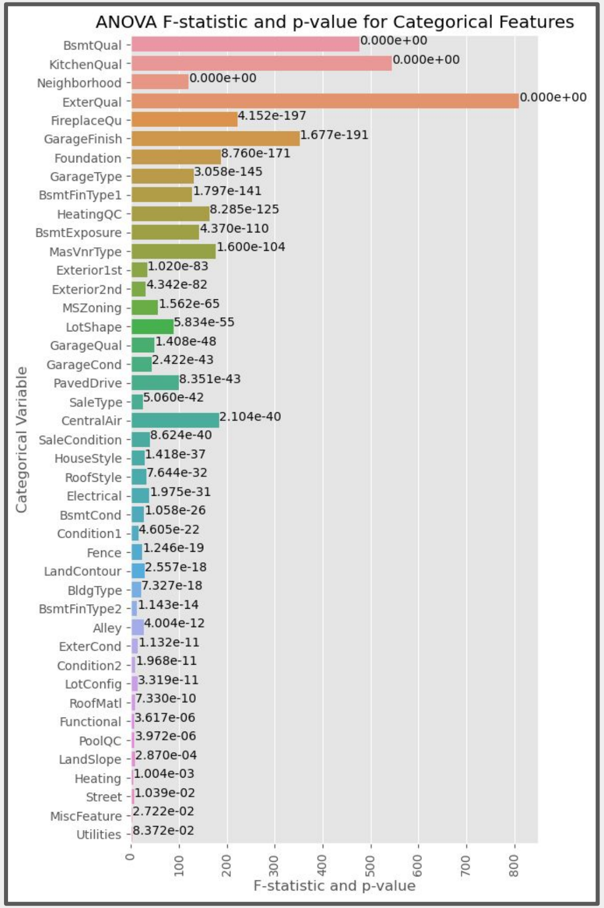
➔ PID, the Parcel Identification Number


Project Workflow
1
Ames housing data
2
Preprocessing
Data cleanup and imputation
Feature engineering and selection
Outlier removal
3
10 preprocessed dataframes
4
Initial modeling to determine best performing dataframe
➔ MLR
➔ Lasso
➔ Ridge
➔ GradientBoostingRegressor
Pipeline:
➔ Scaling numerical variables
➔ Categorical variable encoding
➔ Log transformation of target variable
GridSearchCV with K-fold cross-validation
5
Top performing dataframe
6
Hyperparameter tuning with Optuna
➔ Lasso
➔ Ridge
➔ ElasticNet
➔ XGBoost
➔ GradientBoostingRegressor
➔ RandomForestRegressor
➔ AdaBoostRegressor
Encoding methods
Scaling methods
7
Top performing models
➔ XGboost
➔ ElasticNet
➔ Lasso
8
SHAP

1

EDA: Ames Housing Data





Numerical features highly correlated to Sale Price

➔ OverallQual
➔ GrLivArea

Categorical features with a strong relationship to Sale Price

➔ BsmtQual
➔ KitchenQual
➔ ExterQual
➔ Neighborhood

1

EDA: Ames Housing Data

Large, high quality, inexpensive home

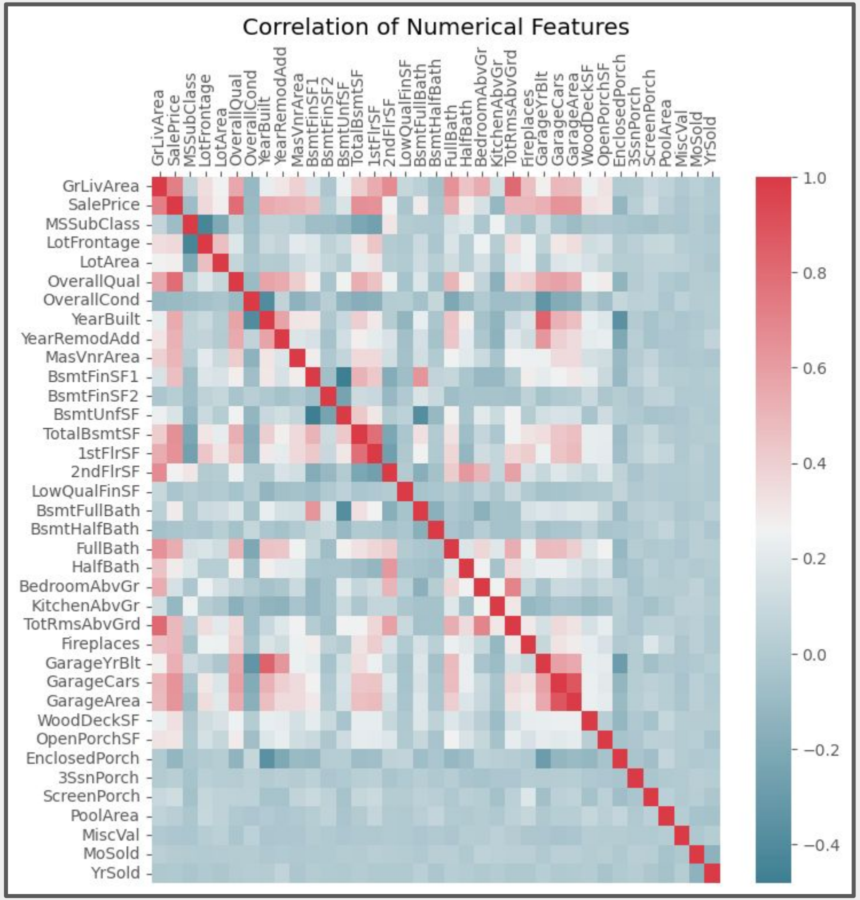
Sale Condition: Partial

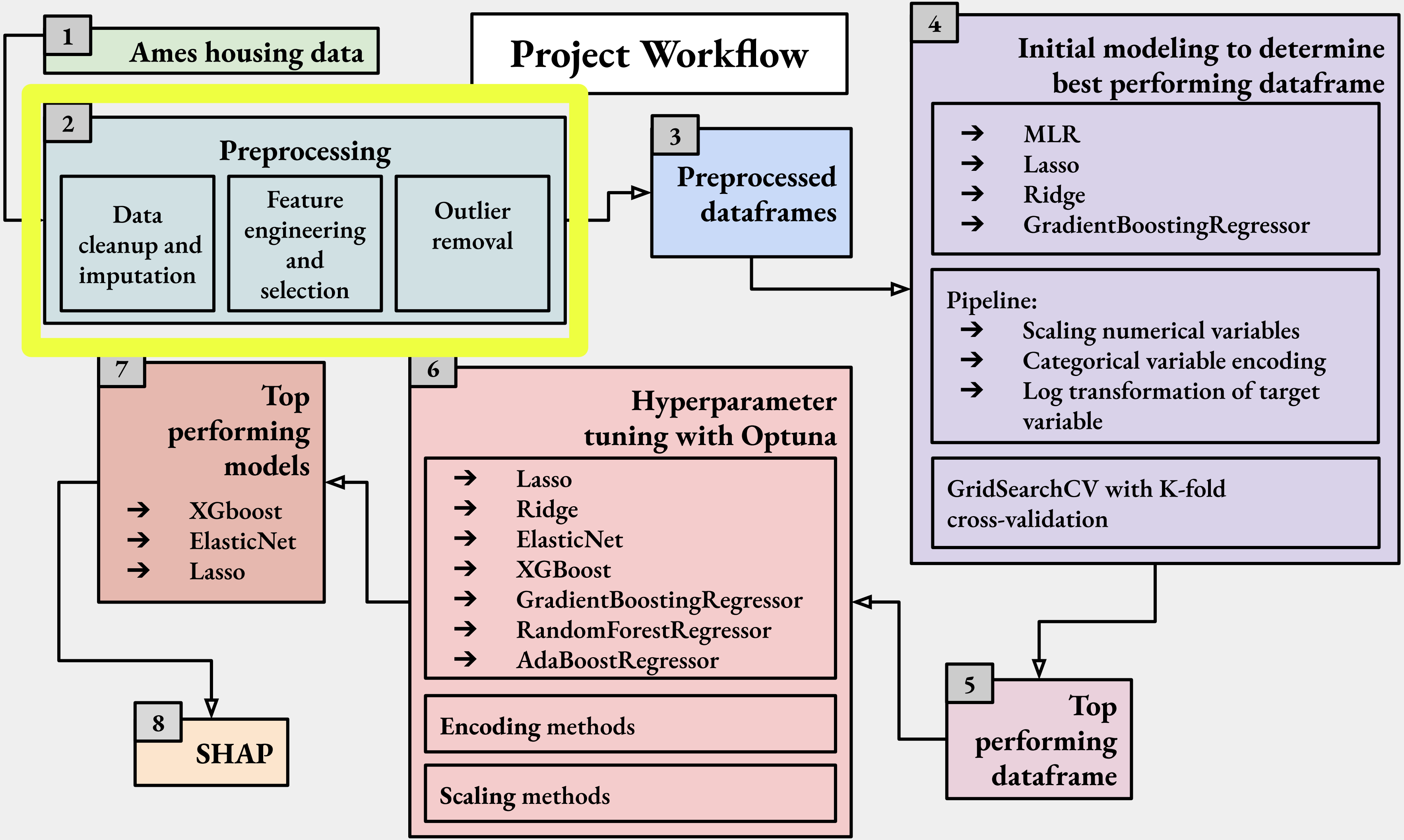
1

EDA: Ames Housing Data

Target variable is positively skewed
➔ Log transformation improves this

Multicollinearity is present among some of the numerical features





Project Workflow
1
Ames housing data
2
Preprocessing
Data cleanup and imputation
Feature engineering and selection
Outlier removal
3
Preprocessed dataframes
4
Initial modeling to determine best performing dataframe
➔ MLR
➔ Lasso
➔ Ridge
➔ GradientBoostingRegressor
Pipeline:
➔ Scaling numerical variables
➔ Categorical variable encoding
➔ Log transformation of target variable
GridSearchCV with K-fold cross-validation
5
Top performing dataframe
6
Hyperparameter tuning with Optuna
➔ Lasso
➔ Ridge
➔ ElasticNet
➔ XGBoost
➔ GradientBoostingRegressor
➔ RandomForestRegressor
➔ AdaBoostRegressor
Encoding methods
Scaling methods
7
Top performing models
➔ XGboost
➔ ElasticNet
➔ Lasso
8
SHAP

2

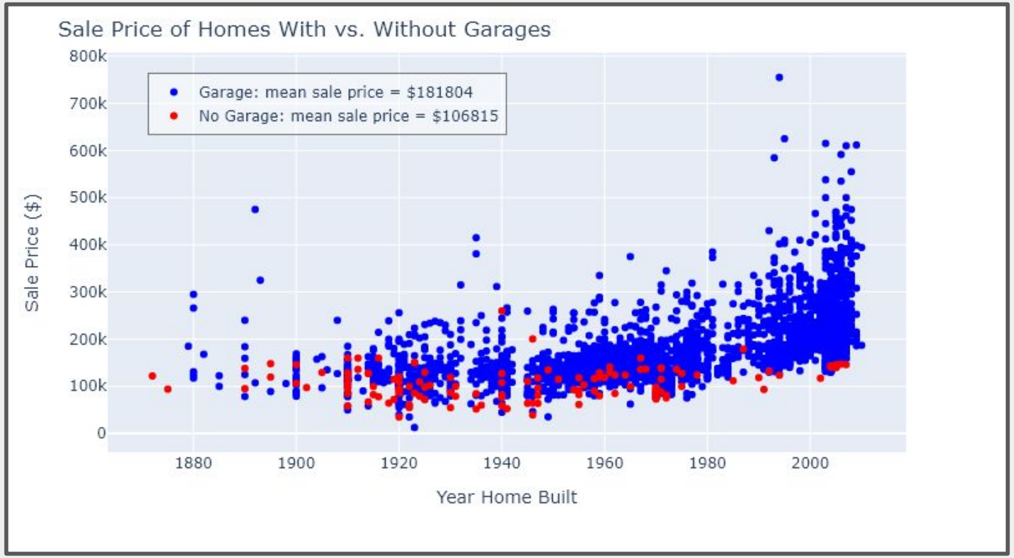
Preprocessing

- → Convert MSSubClass, MoSold, YrSold from discrete numerical variables to nominal categorical variables.
- → Convert ExterQual, ExterCond, KitchenQual, BsmtQual, BsmtCond, and many others from ordinal categorical to discrete numerical variables.

2

Preprocessing

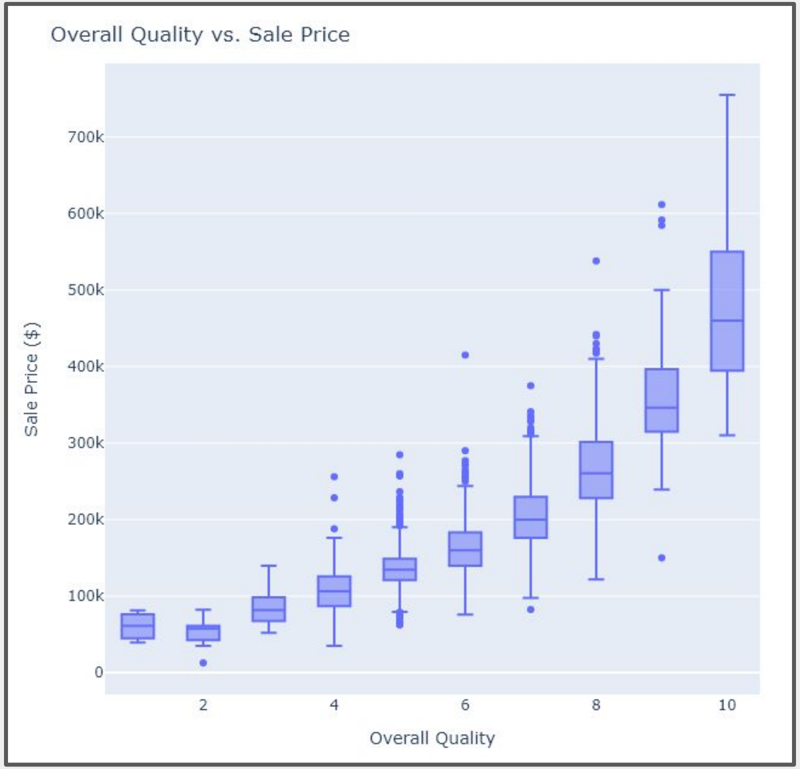
- ➔ Most numerical nulls filled with 0 and most categorical nulls filled with 'NO'
- ➔ LotFrontage nulls changed to a percentage of the lot area based on the mean percent of lot area
- ➔ PoolArea and GarageYrBlt changed to 'yes' or 'no' categorical variables



2

Preprocessing

Overall Quality was squared & Kitchen Quality and External Quality were cubed in some dataframes

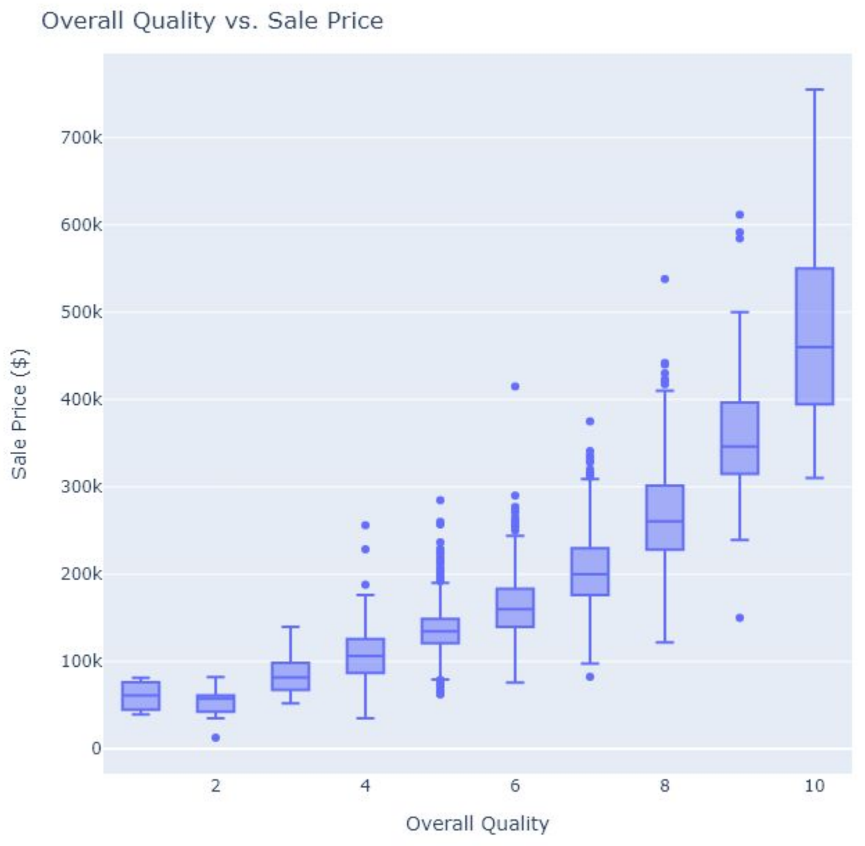


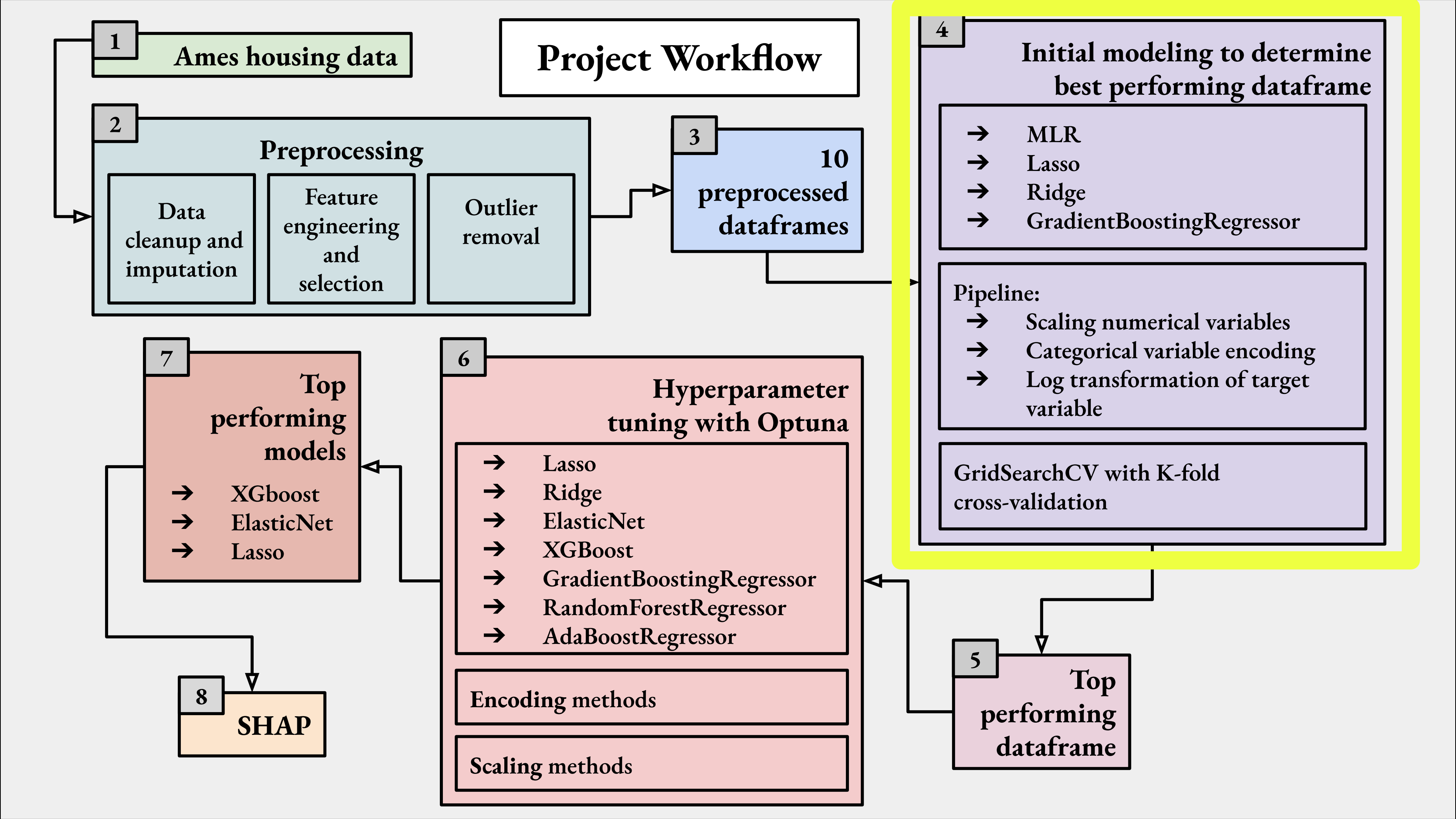
3

Preprocessed Dataframes

8 dataframes with different methods of outlier removal and feature scaling

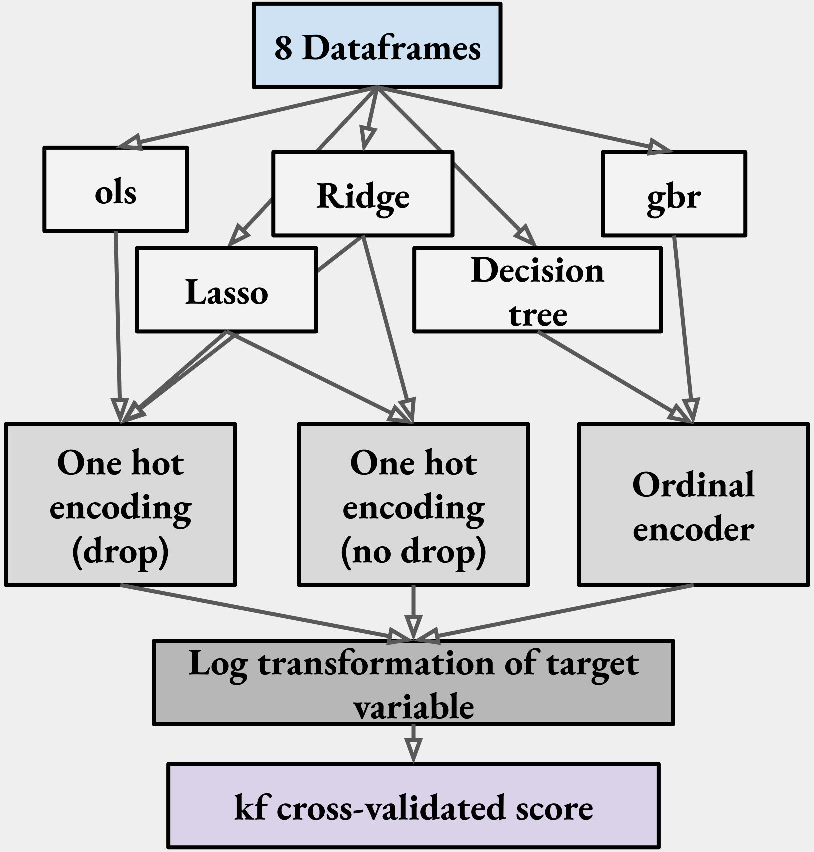
1. 4676 square foot home removed
2. All outliers removed
3. Non-normal sale removed
4. Outliers with quality groups removed
5. Non-normal sales & outliers within quality groups removed
6. Quality features unscaled/4676 square foot home removed
7. Quality features unscaled/Non-normal sale removed
8. Quality features unscaled/Non-normal sales & outliers within quality groups removed



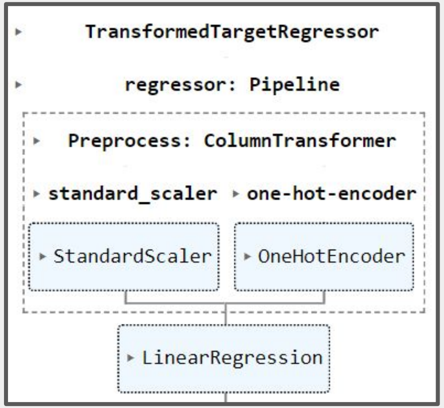

Project Workflow
1
Ames housing data
2
Preprocessing
Data cleanup and imputation
Feature engineering and selection
Outlier removal
3
10 preprocessed dataframes
4
Initial modeling to determine best performing dataframe
MLR
Lasso
Ridge
GradientBoostingRegressor
Pipeline:
Scaling numerical variables
Categorical variable encoding
Log transformation of target variable
GridSearchCV with K-fold cross-validation
5
Top performing dataframe
6
Hyperparameter tuning with Optuna
Lasso
Ridge
ElasticNet
XGBoost
GradientBoostingRegressor
RandomForestRegressor
AdaBoostRegressor
Encoding methods
Scaling methods
7
Top performing models
XGboost
ElasticNet
Lasso
8
SHAP

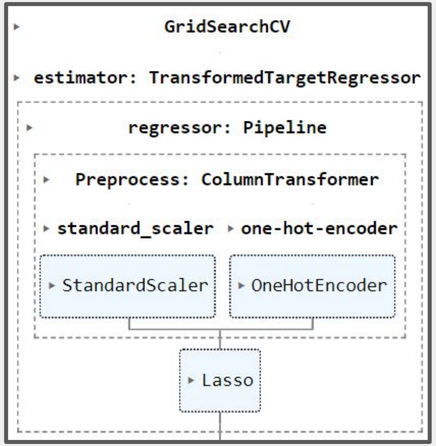
4

Initial Modeling



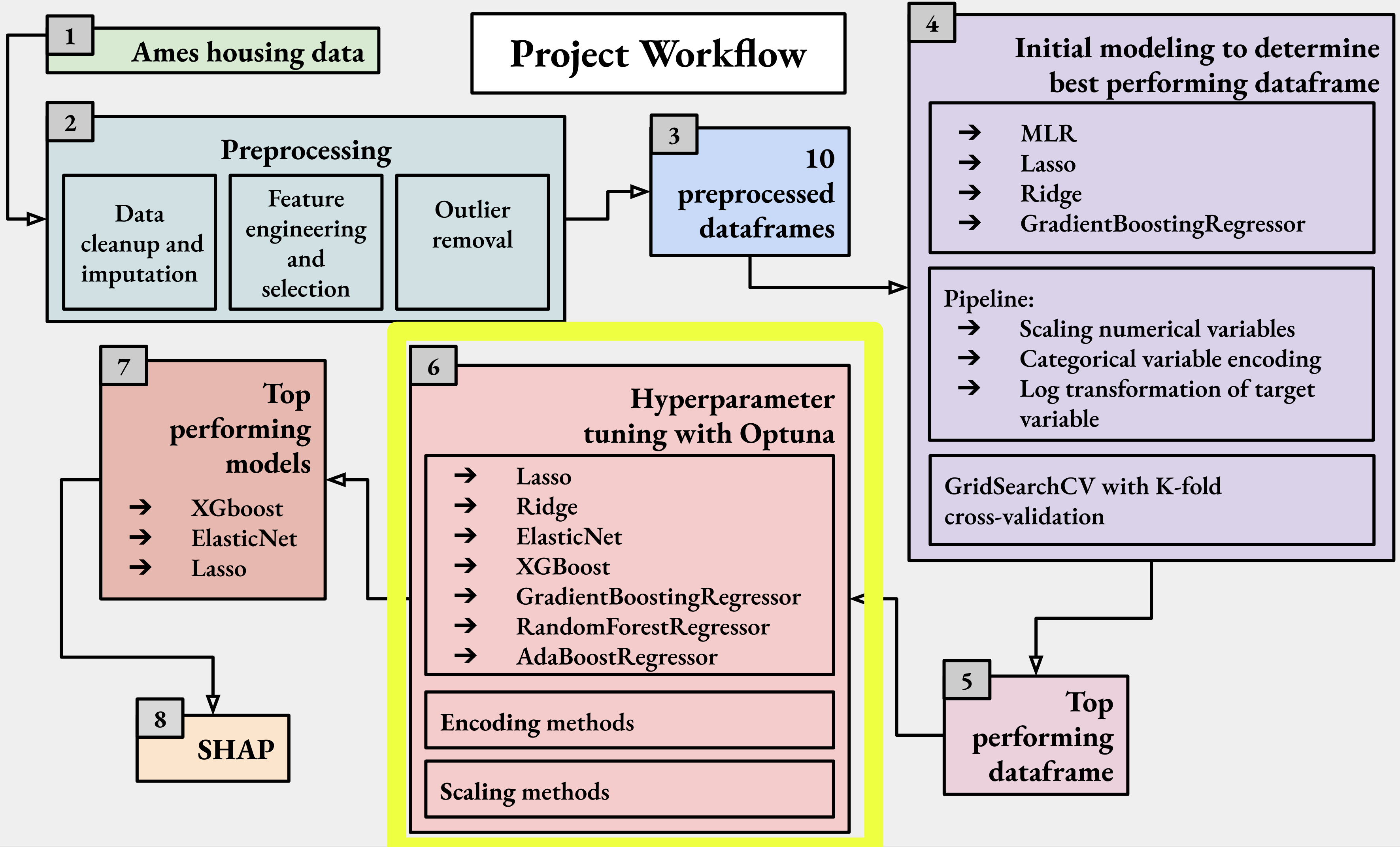
- ➔ “Data leakage occurs when information that would not be available at prediction time is used when building the model.” -sklearn
- ➔ Using a pipeline with K-Fold cross-validator prevents any data leakage when scaling features or transforming the target variable





5 Top Performing Dataframe

Observations Removed	ols	ridge	ridge (drop)	lasso	lasso (drop)	decision tree	gbr
all outliers	0.90882	0.91776	0.91742	0.91735	0.91768	0.75738	0.91380
non-normal sales	0.92017	0.93098	0.93023	0.93316	0.93294	0.80043	0.93053
non-normal sales/quality-group outliers	**0.94577**	**0.94969**	**0.94953**	**0.94967**	**0.94963**	0.81488	0.93578
quality-group outliers	**0.94254**	**0.94672**	**0.94652**	**0.94602**	**0.94615**	0.81107	0.93053
4676 square foot home	0.92372	0.93029	0.92957	0.93027	0.93017	0.77206	0.92773
unscaled/non-normal sales	0.91940	0.92981	0.92885	0.93092	0.93075	0.80458	0.93074
unscaled/non-normal sales/quality-group outliers	0.92463	0.93194	0.93211	0.93394	0.93386	0.76464	0.93109
unscaled/4676 square foot home	0.92235	0.92957	0.92880	0.92953	0.92942	0.79030	0.92798

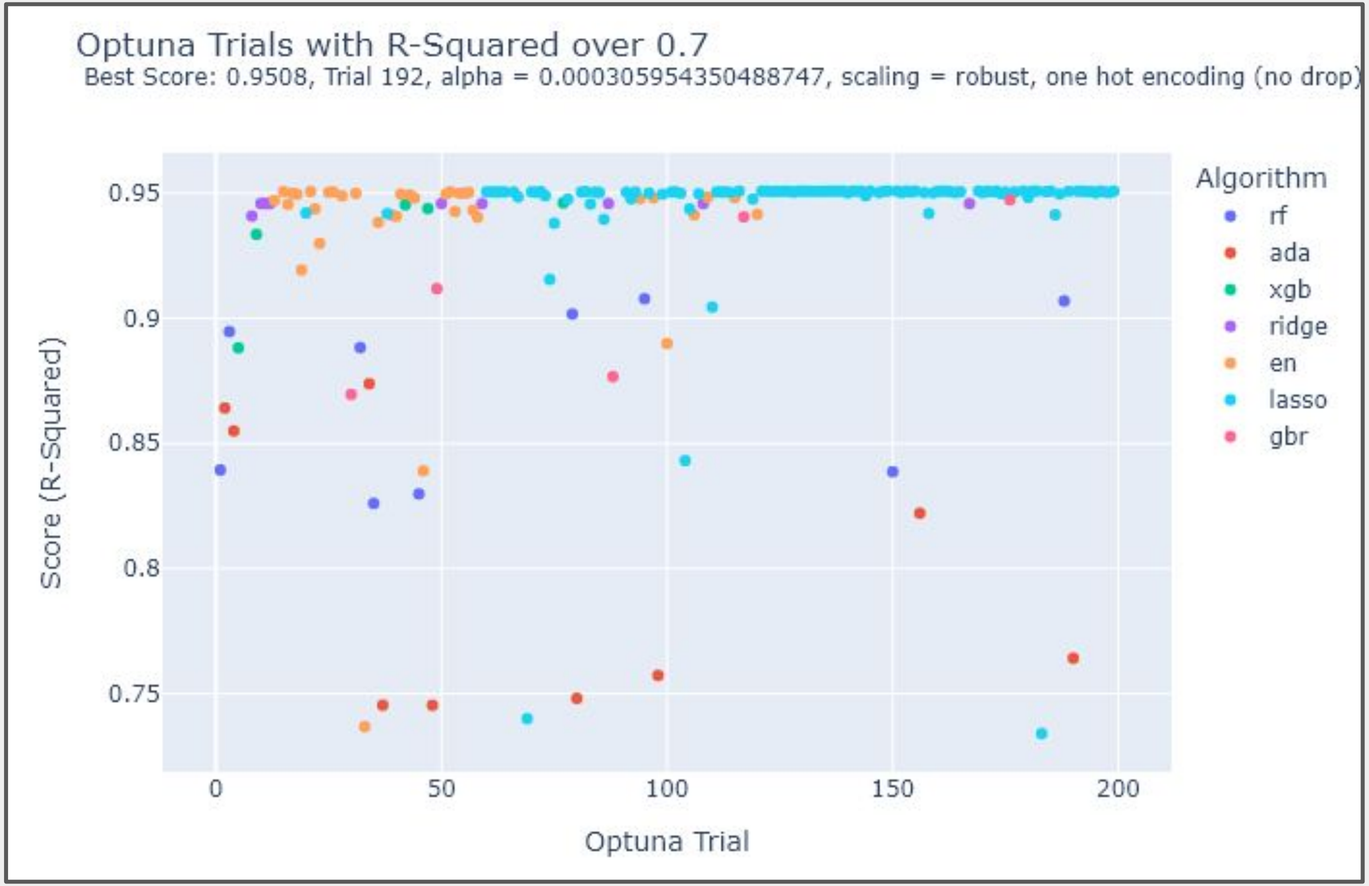

Project Workflow
1
Ames housing data
2
Preprocessing
Data cleanup and imputation
Feature engineering and selection
Outlier removal
3
10 preprocessed dataframes
4
Initial modeling to determine best performing dataframe
➔ MLR
➔ Lasso
➔ Ridge
➔ GradientBoostingRegressor
Pipeline:
➔ Scaling numerical variables
➔ Categorical variable encoding
➔ Log transformation of target variable
GridSearchCV with K-fold cross-validation
5
Top performing dataframe
6
Hyperparameter tuning with Optuna
➔ Lasso
➔ Ridge
➔ ElasticNet
➔ XGBoost
➔ GradientBoostingRegressor
➔ RandomForestRegressor
➔ AdaBoostRegressor
Encoding methods
Scaling methods
7
Top performing models
➔ XGboost
➔ ElasticNet
➔ Lasso
8
SHAP

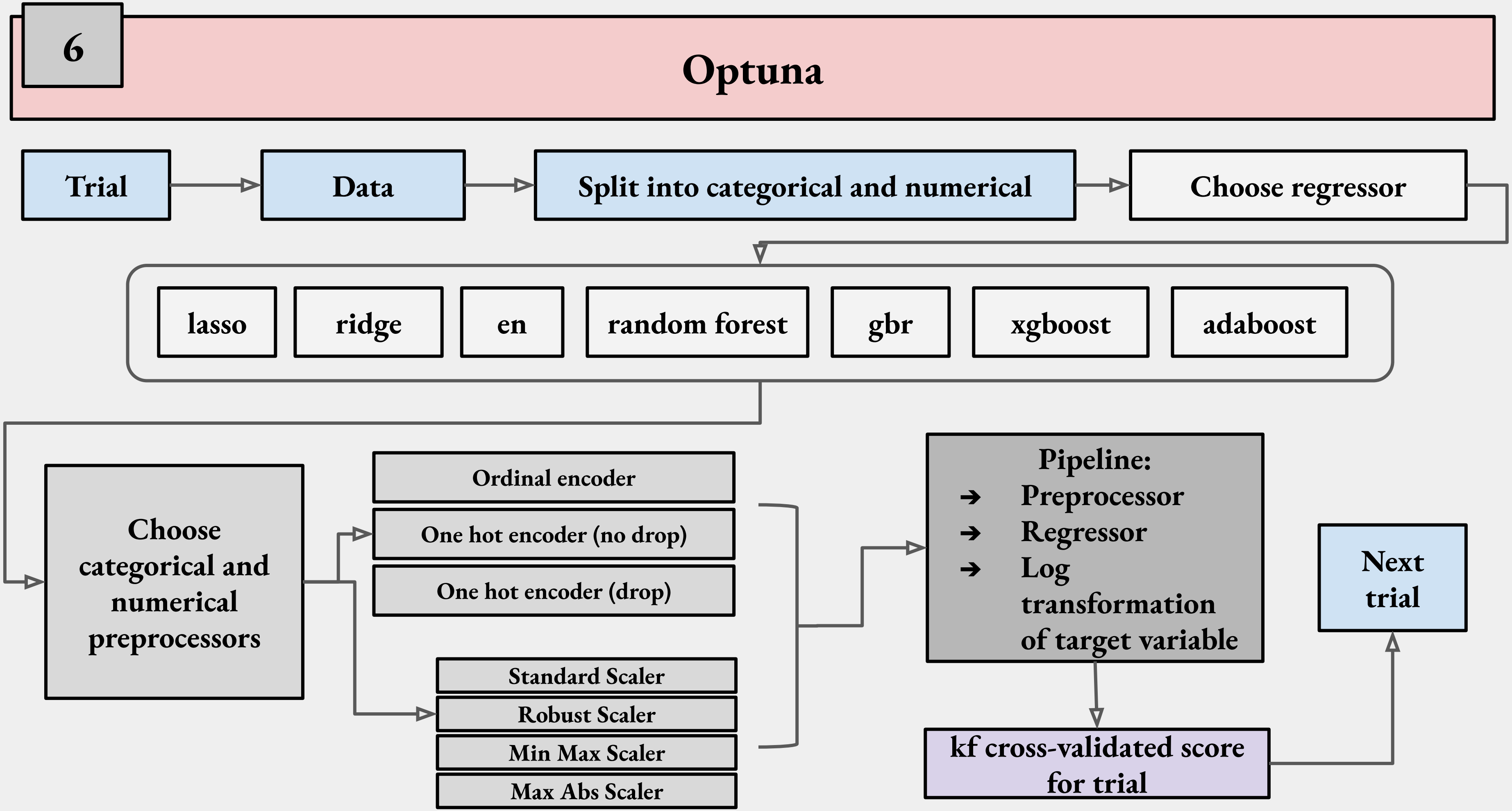
Optuna

Optuna is an open source hyperparameter optimization framework

Hyperparameters:

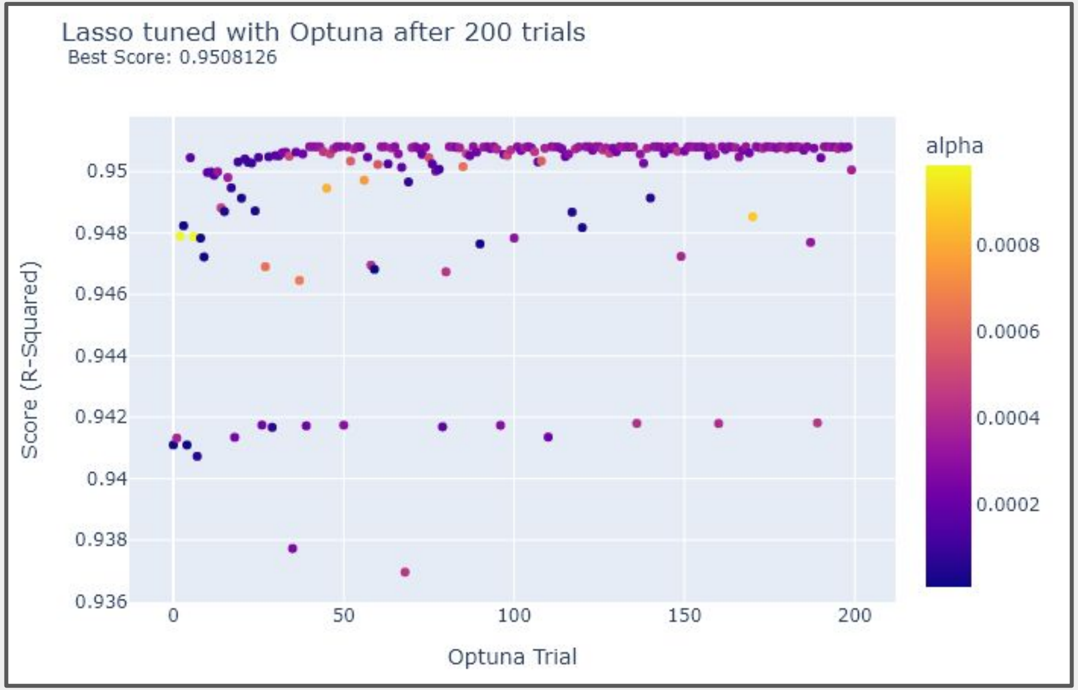
- ➔ Hyperparameters for the algorithm
- ➔ Algorithm
- ➔ Encoding method
- ➔ Scaling method

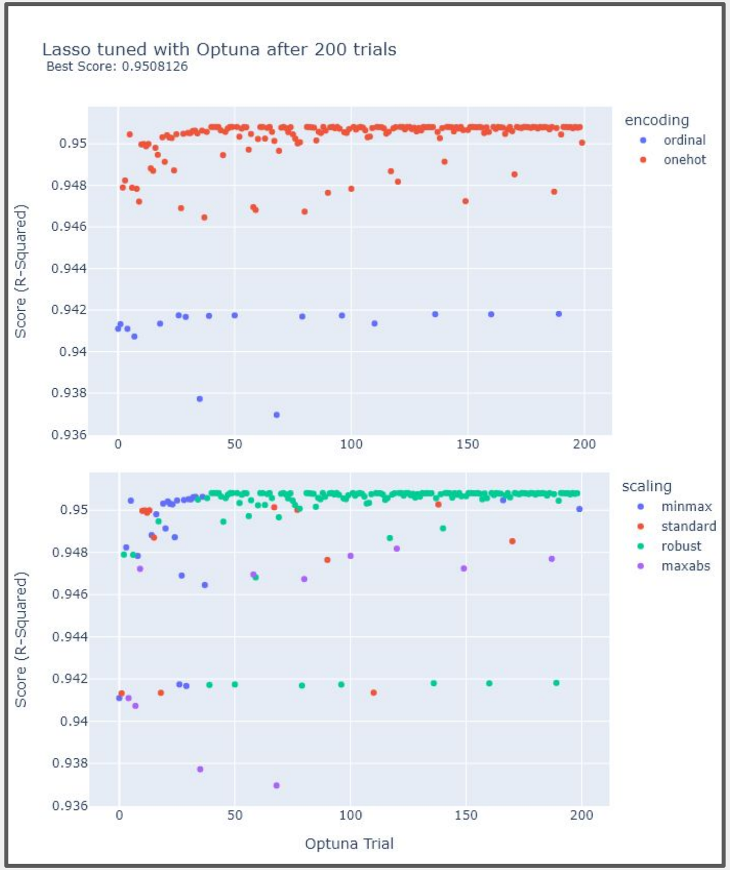



6
Optuna
Trial
Data
Split into categorical and numerical
Choose regressor
lasso
ridge
en
random forest
gbr
xgboost
adaboost
Choose categorical and numerical preprocessors
Ordinal encoder
One hot encoder (no drop)
One hot encoder (drop)
Standard Scaler
Robust Scaler
Min Max Scaler
Max Abs Scaler
Pipeline:
➔ Preprocessor
➔ Regressor
➔ Log transformation of target variable
Next trial
kf cross-validated score for trial

6 Optuna - Lasso

0.95081254796895
{'scaling_method': 'robust', 'encoding_method': 'onehot', 'alpha': 0.0003059499593517016}

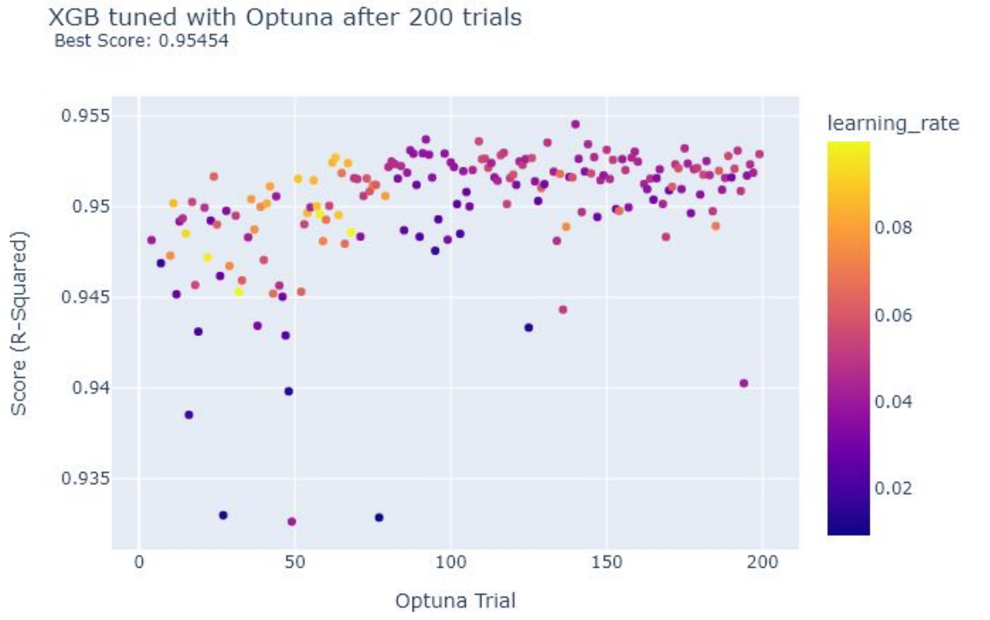


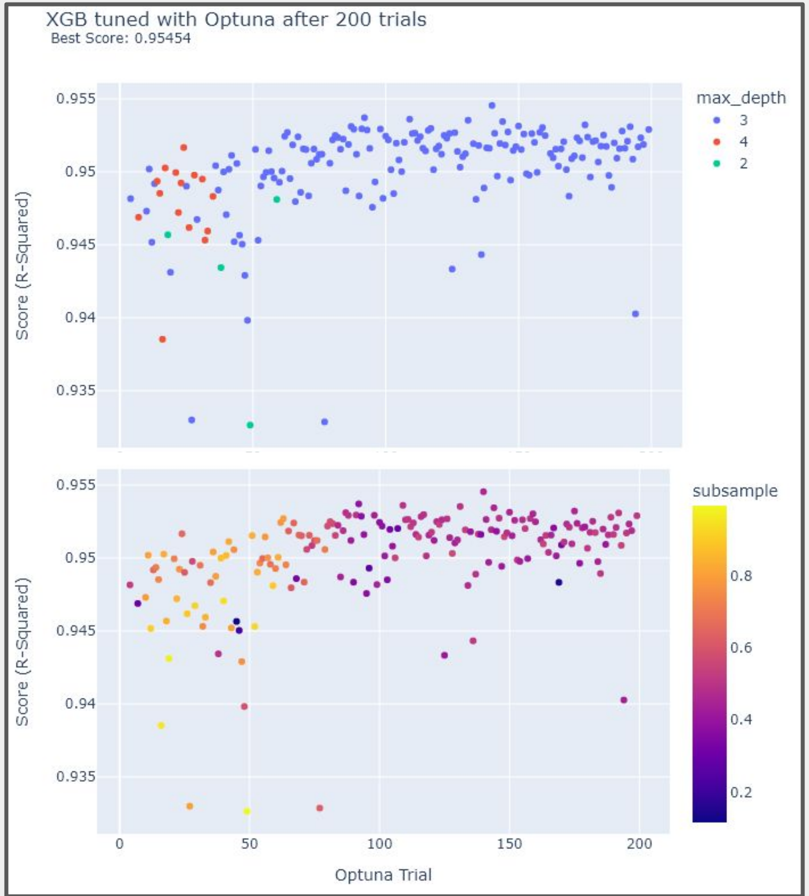


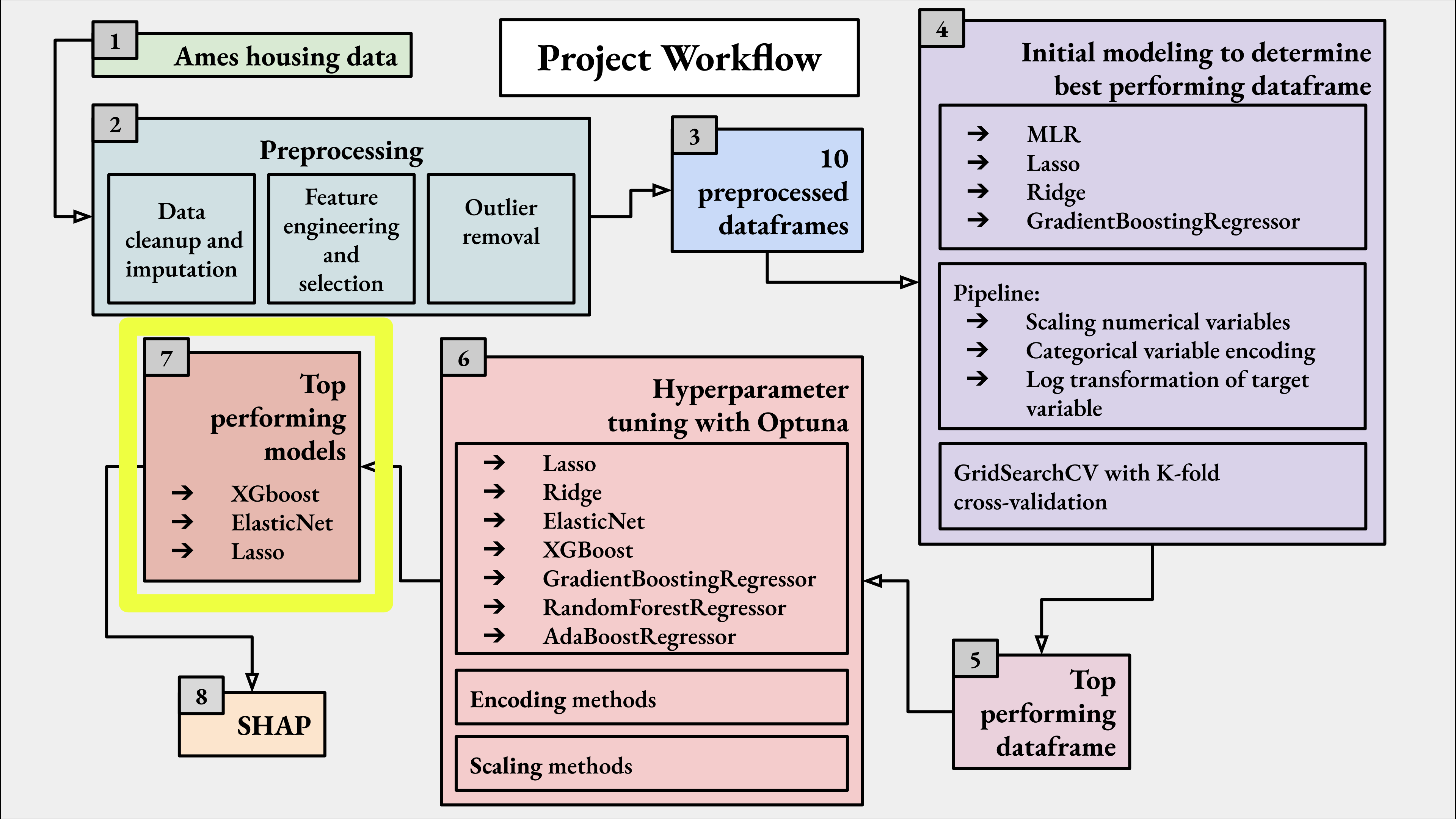
6

Optuna - XGBoost

0.9545422235903054
{'scaling_method': 'standard', 'encoding_method': 'onehot', 'n_estimators': 902, 'learning_rate': 0.04089478271640344, 'max_depth': 3, 'subsample': 0.4614417149387252, 'colsample_bytree': 0.6589253772701361, 'min_child_weight': 2}

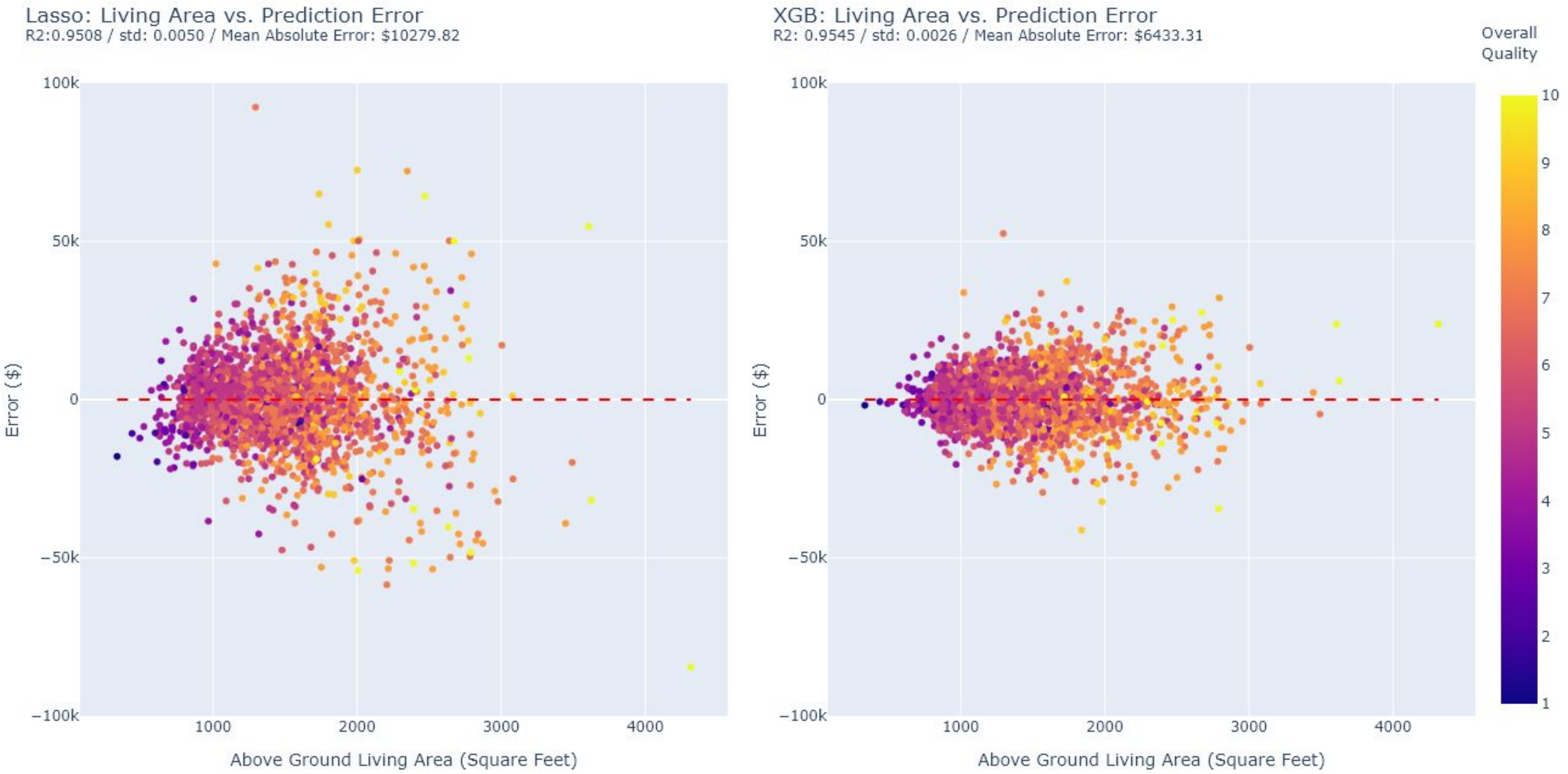


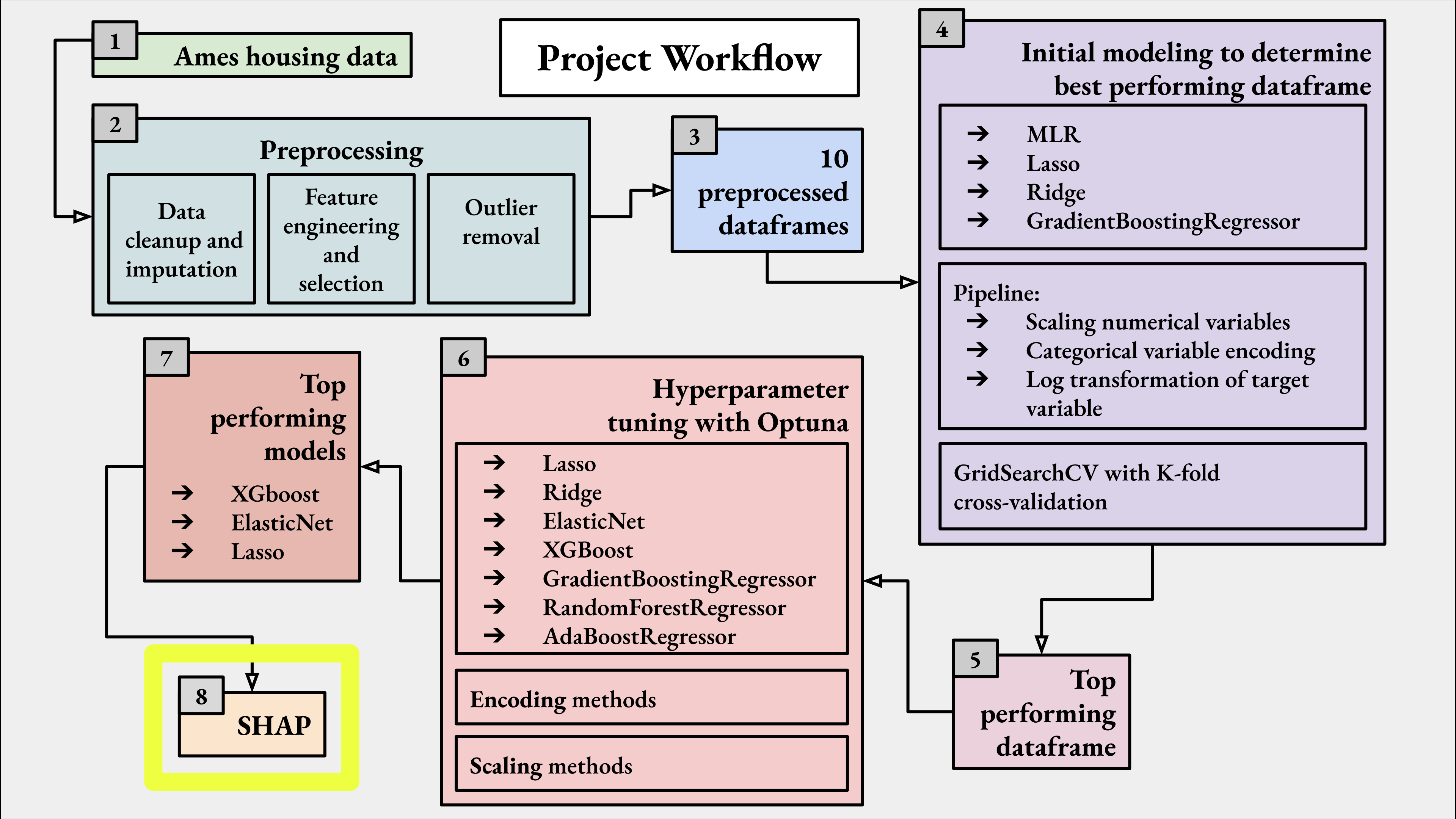



Project Workflow
1
Ames housing data
2
Preprocessing
Data cleanup and imputation
Feature engineering and selection
Outlier removal
3
10 preprocessed dataframes
4
Initial modeling to determine best performing dataframe
➔ MLR
➔ Lasso
➔ Ridge
➔ GradientBoostingRegressor
Pipeline:
➔ Scaling numerical variables
➔ Categorical variable encoding
➔ Log transformation of target variable
GridSearchCV with K-fold cross-validation
5
Top performing dataframe
6
Hyperparameter tuning with Optuna
➔ Lasso
➔ Ridge
➔ ElasticNet
➔ XGBoost
➔ GradientBoostingRegressor
➔ RandomForestRegressor
➔ AdaBoostRegressor
Encoding methods
Scaling methods
7
Top performing models
➔ XGboost
➔ ElasticNet
➔ Lasso
8
SHAP

7

Top Performing Models




Project Workflow
1
Ames housing data
2
Preprocessing
Data cleanup and imputation
Feature engineering and selection
Outlier removal
3
10 preprocessed dataframes
4
Initial modeling to determine best performing dataframe
➔ MLR
➔ Lasso
➔ Ridge
➔ GradientBoostingRegressor
Pipeline:
➔ Scaling numerical variables
➔ Categorical variable encoding
➔ Log transformation of target variable
GridSearchCV with K-fold cross-validation
5
Top performing dataframe
6
Hyperparameter tuning with Optuna
➔ Lasso
➔ Ridge
➔ ElasticNet
➔ XGBoost
➔ GradientBoostingRegressor
➔ RandomForestRegressor
➔ AdaBoostRegressor
Encoding methods
Scaling methods
7
Top performing models
➔ XGboost
➔ ElasticNet
➔ Lasso
8
SHAP

SHAP

"SHAP (SHapley Additive exPlanations) is a game theoretic approach to explain the output of any machine learning model. It connects optimal credit allocation with local explanations using the classic Shapley values from game theory and their related extensions (see papers for details and citations)." -SHAP

Game Theory: branch of mathematics concerned with the analysis of strategies for dealing with competitive situations where the outcome of a participant's choice of action depends critically on the actions of other participants.

➔ The Shapley value of a feature value is not the difference of the predicted value after removing the feature from the model training. The interpretation of the Shapley value is: Given the current set of feature values, the contribution of a feature value to the difference between the actual prediction and the mean prediction is the estimated Shapley value.

https://christophm.github.io/interpretable-ml-book/shapley.html

8

SHAP

"Players? Game? Payout? What is the connection to machine learning predictions and interpretability?

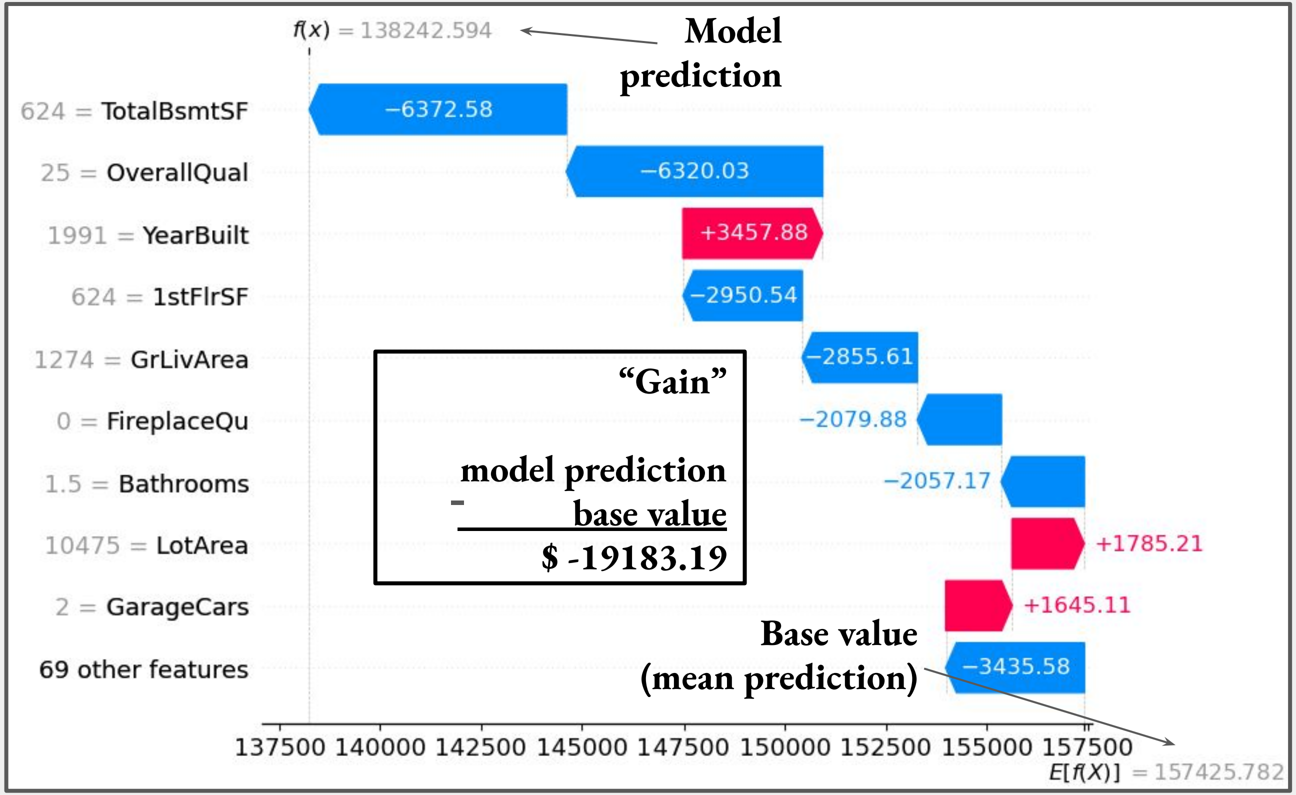
The "game" is the prediction task for a single instance of the dataset.

The "gain" is the actual prediction for this instance minus the average prediction for all instances.

The "players" are the feature values of the instance that collaborate to receive the gain (= predict a certain value)."

https://christophm.github.io/interpretable-ml-book/shapley.html

Example:

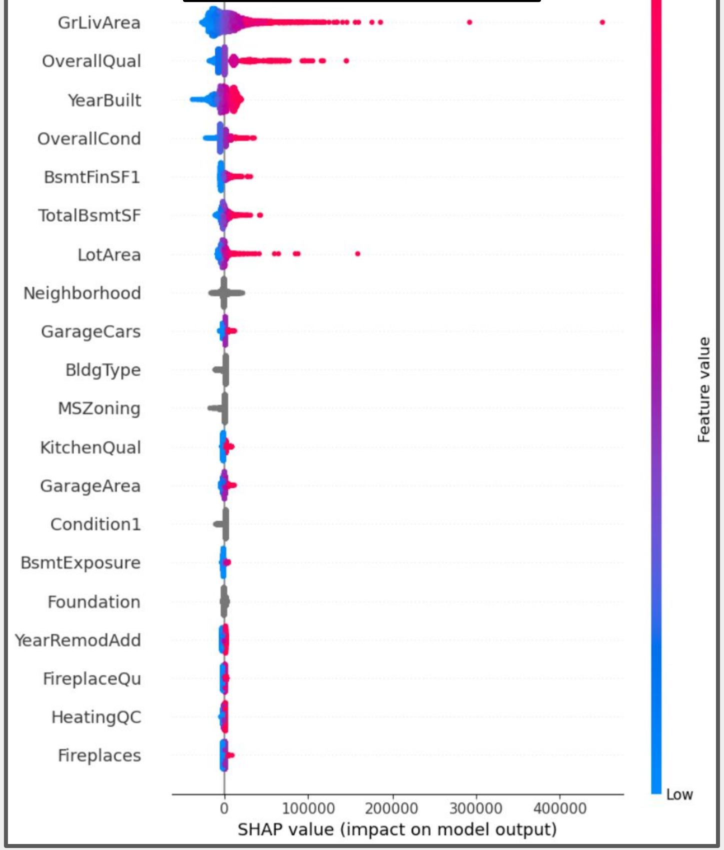


8

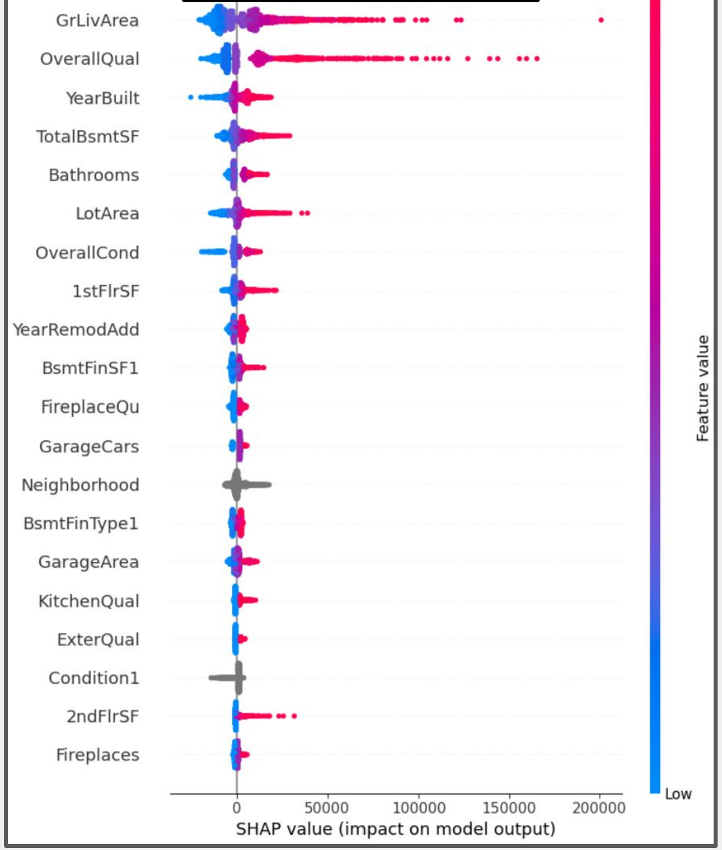
SHAP - lasso/xgb

SHAP summary plots show more nuance in the XGBoost model

Summary Plot - Lasso

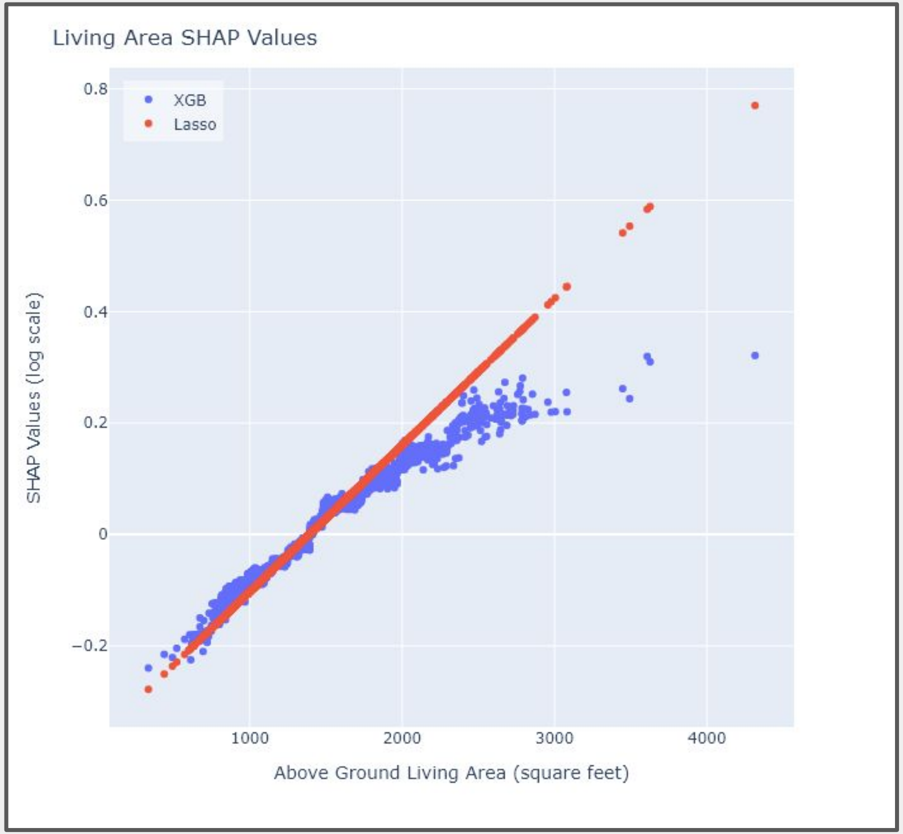


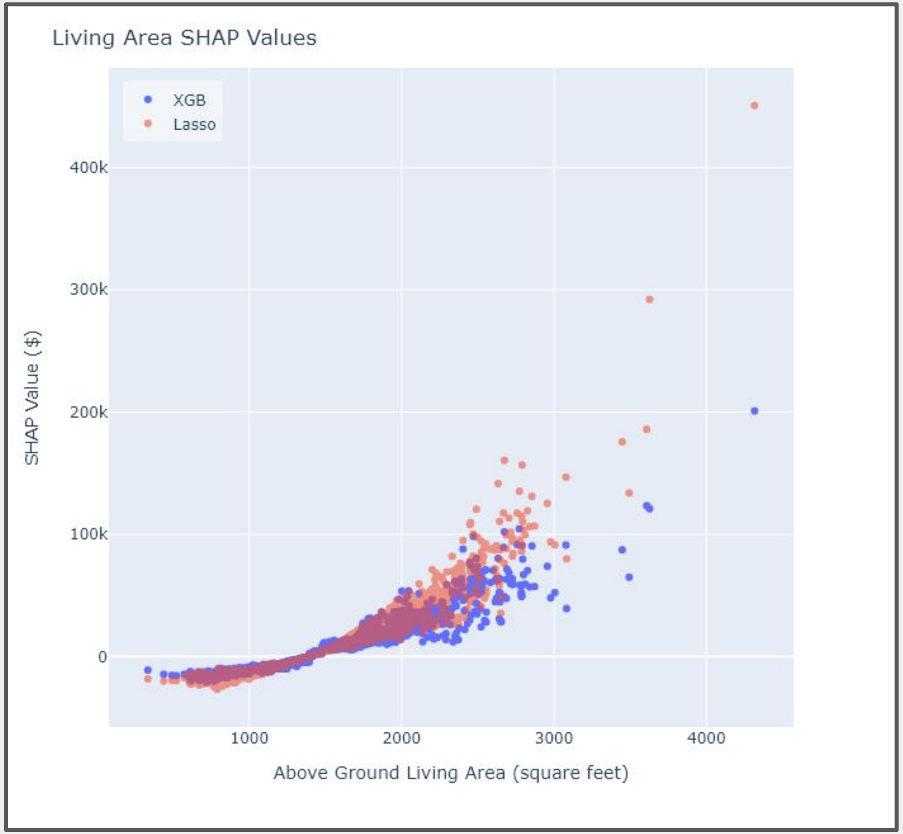
Summary Plot - XGBoost



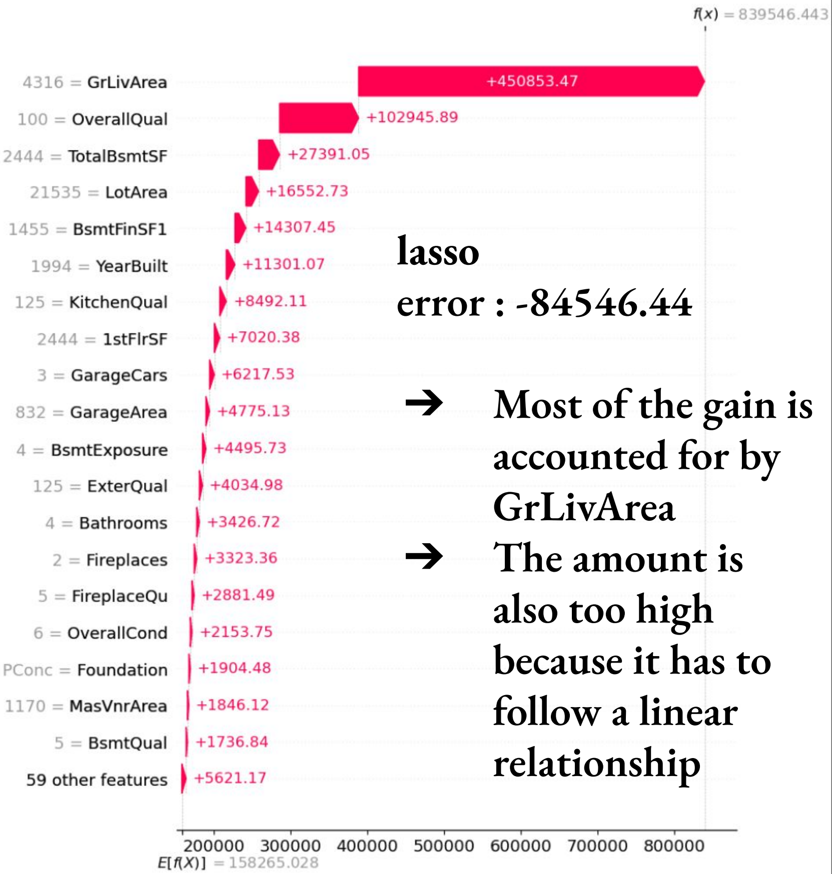
8

SHAP - Above Ground Living Area





SHAP - GrLivArea: 4316/ SalePrice: $ 755,000

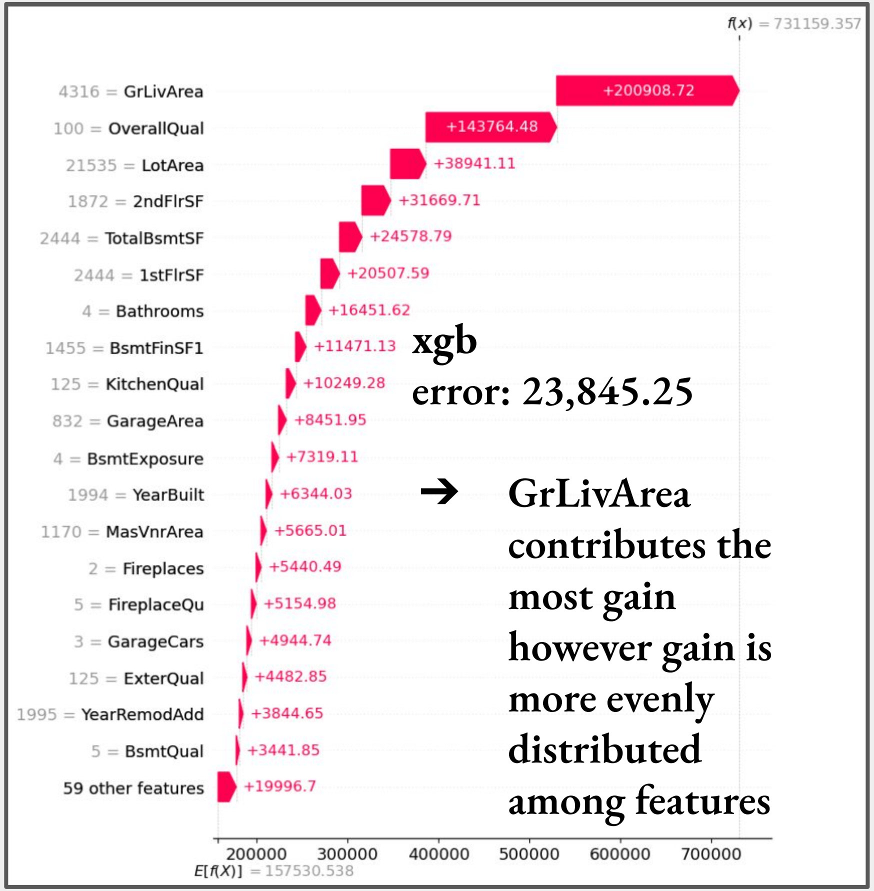


lasso

error : -84546.44

➔ Most of the gain is accounted for by GrLivArea

➔ The amount is also too high because it has to follow a linear relationship

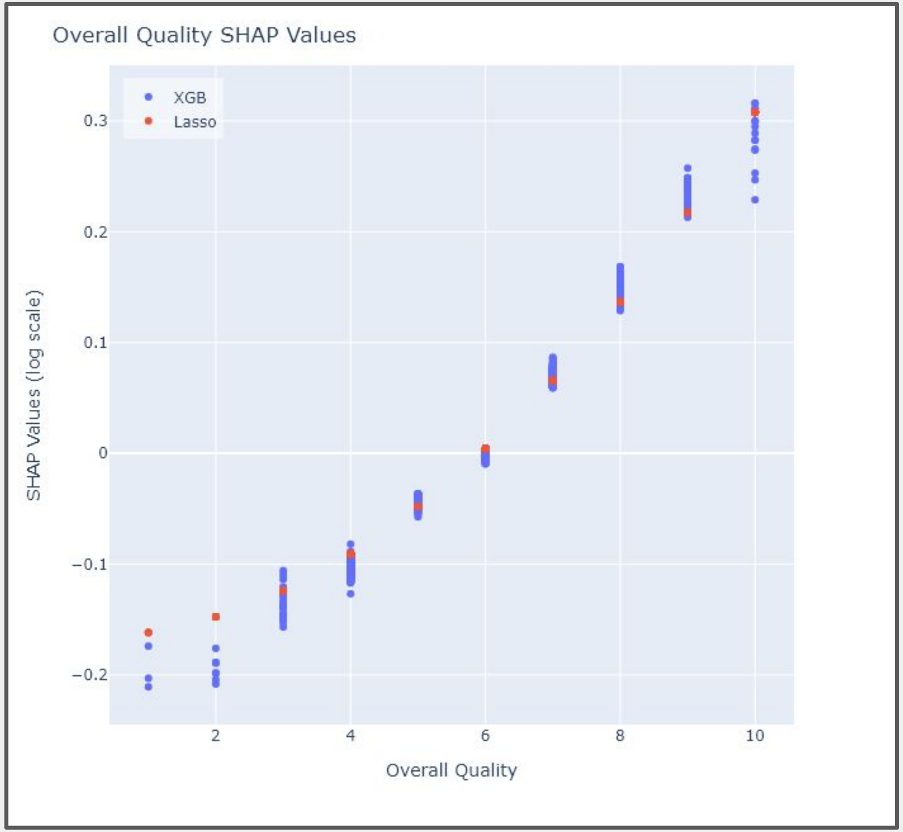


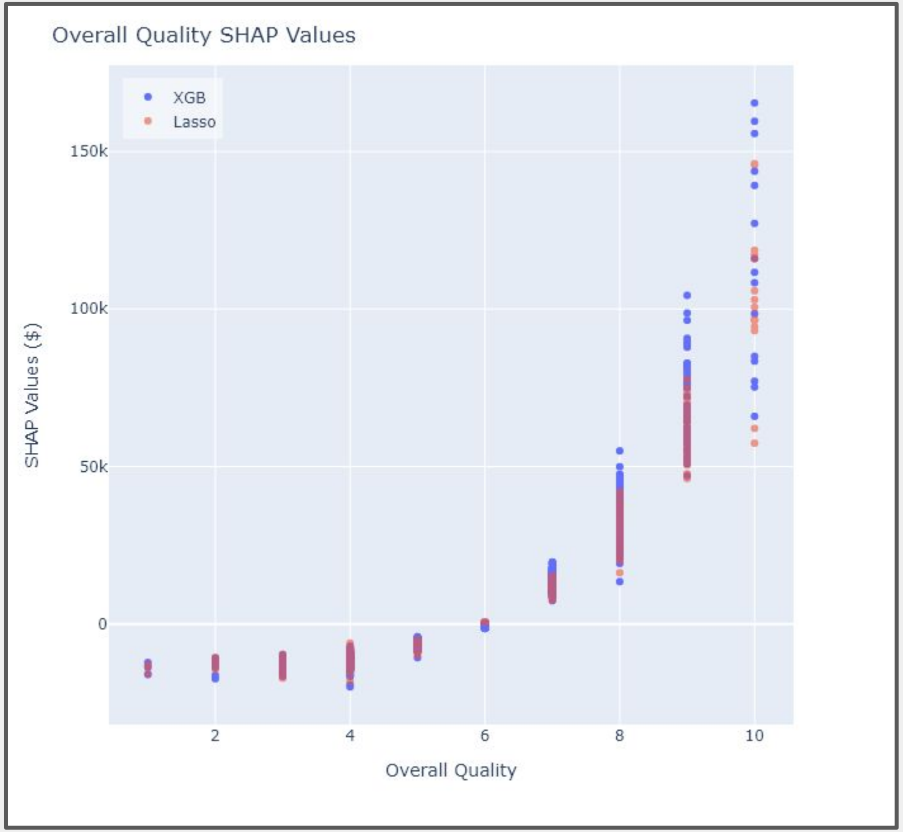
xgb

error: 23,845.25

➔ GrLivArea contributes the most gain however gain is more evenly distributed among features

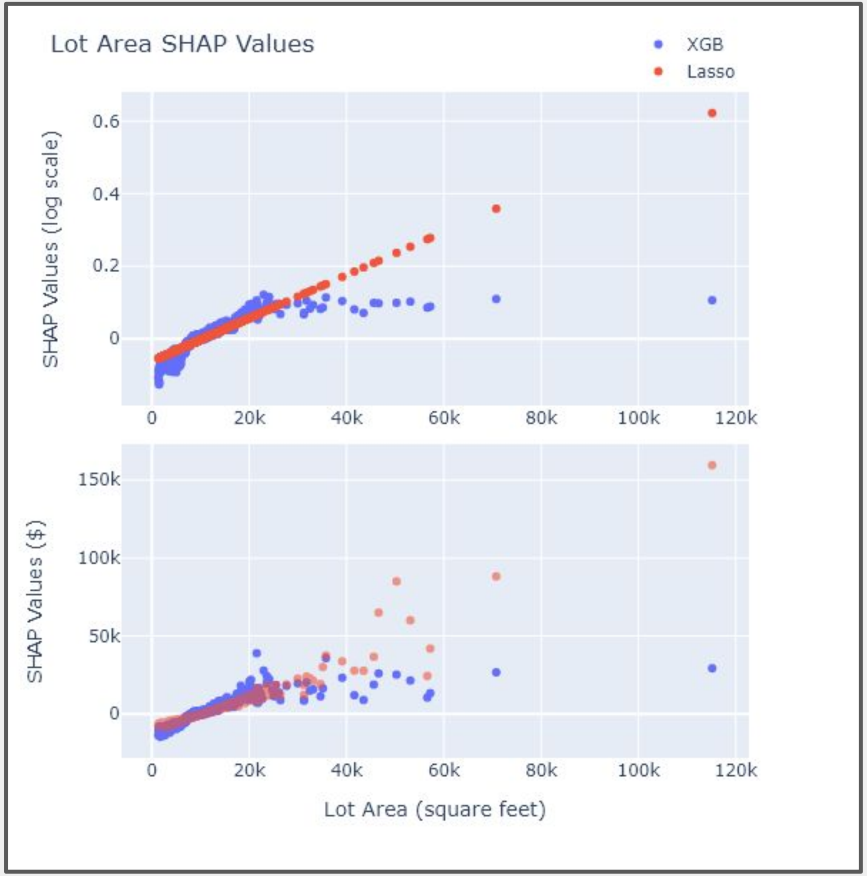
SHAP - Overall Quality





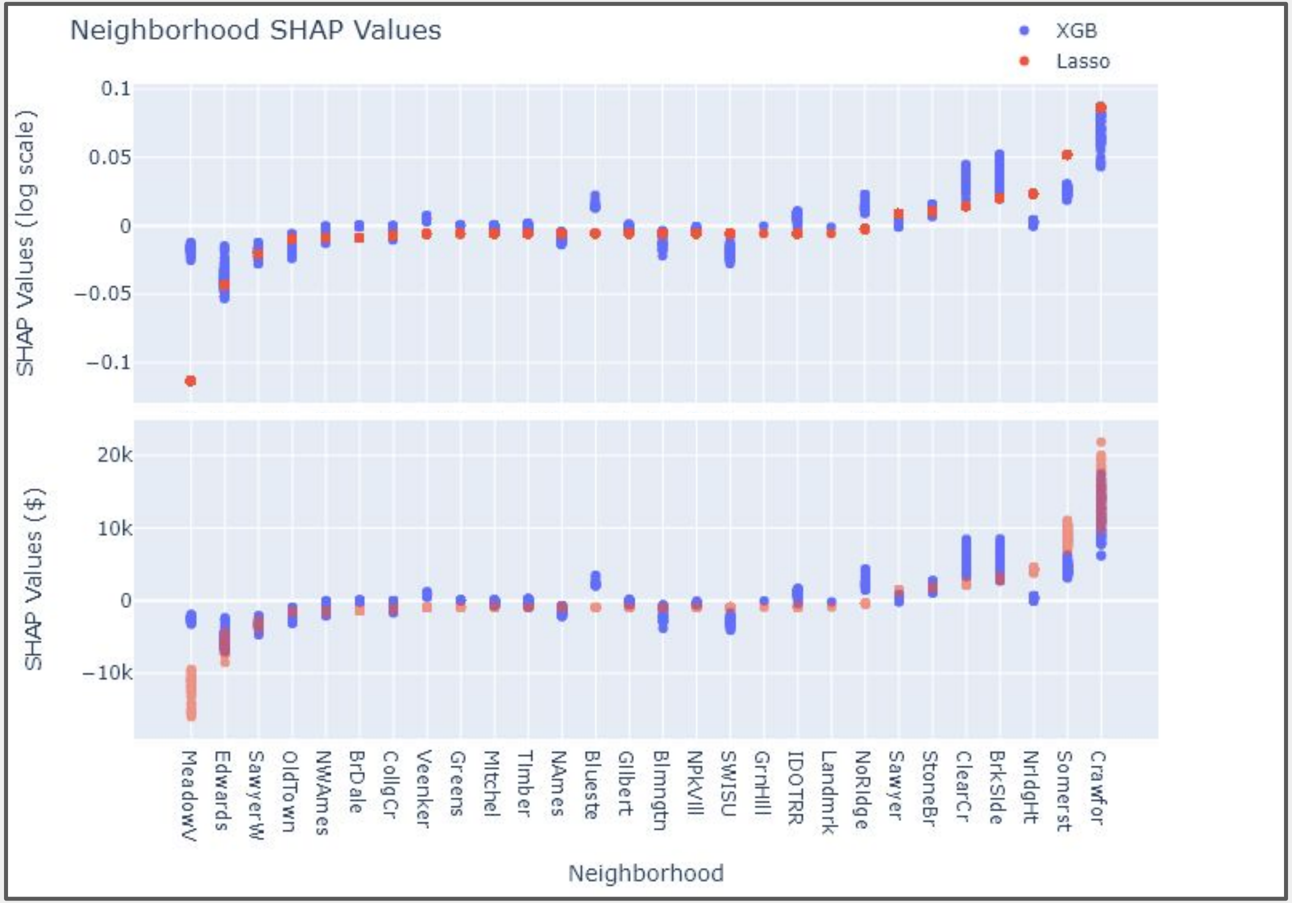
8

SHAP - Lot Area/Garage Cars



8

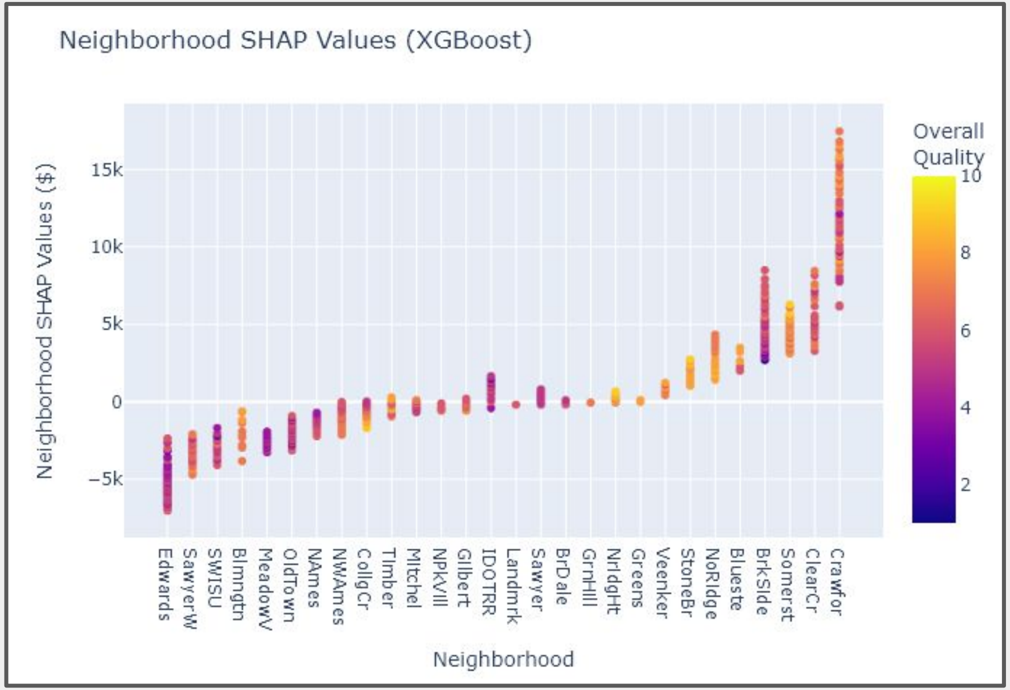
SHAP - Neighborhood



8

SHAP - Neighborhood (XGBoost)

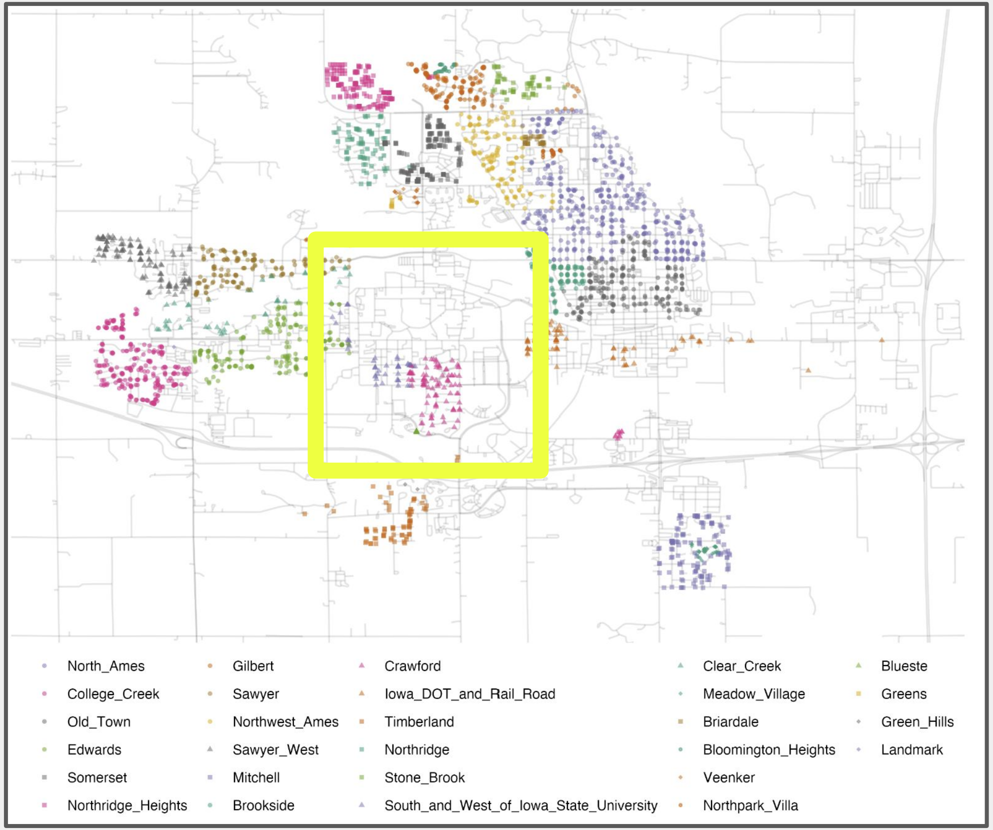
Clear Creek, Brookside, and Crawford have high neighborhood SHAP values relative to many homes with higher Overall Quality. Within their quality groups they also have the highest neighborhood SHAP values.



SHAP - Crawford vs. SWISU

Both the lasso and xgb models have high SHAP values for Crawford and negative SHAP values for SWISU

- ➔ Both close to campus
- ➔ Similar size homes
- ➔ Similar quality
- ➔ Crawford homes have larger lots and are on average about 10 years newer
- ➔ Crawford homes sell for an average ~ $60000 more

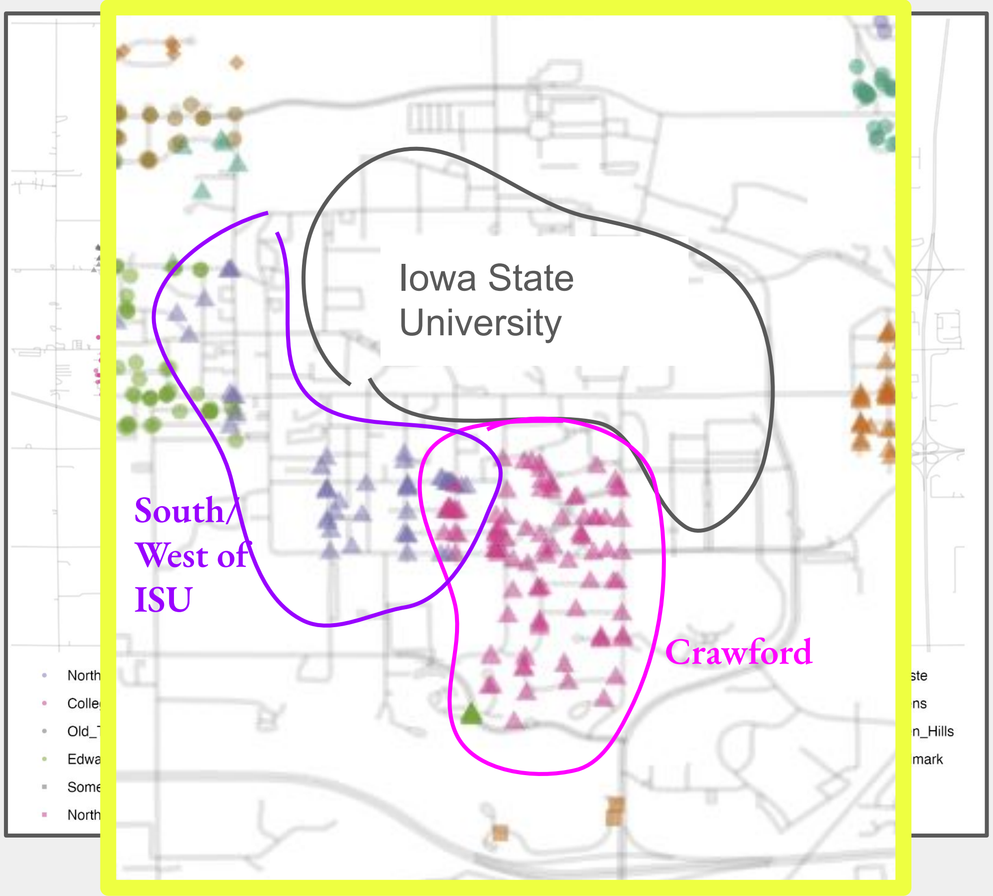


https://www.tmwr.org/ames

SHAP - Crawford vs. SWISU

Both the lasso and xgb models have high SHAP values for Crawford and negative SHAP values for SWISU

- ➔ Both close to campus
- ➔ Similar size homes
- ➔ Similar quality
- ➔ Crawford homes have larger lots and are on average about 10 years newer
- ➔ Crawford homes sell for an average ~ $60000 more

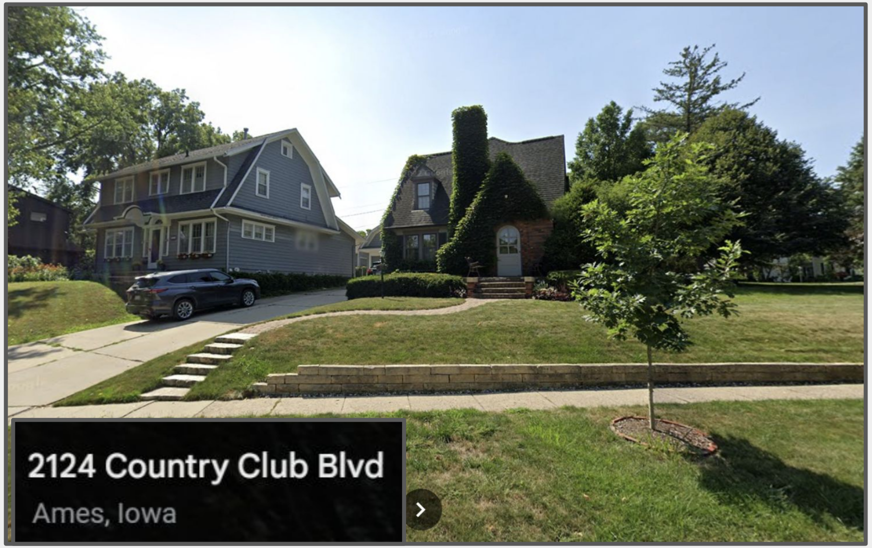


https://www.tmwr.org/ames

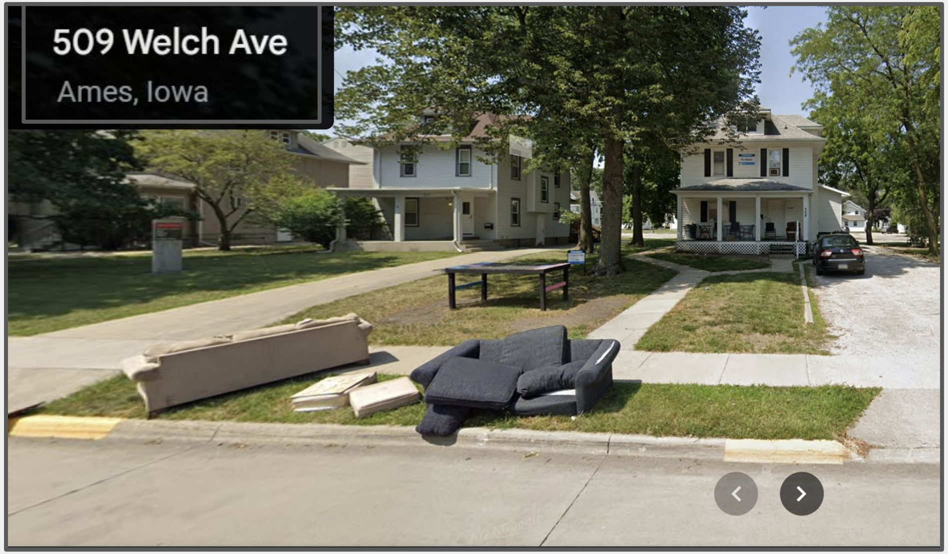
8

SHAP - Crawford vs. SWISU

Crawford

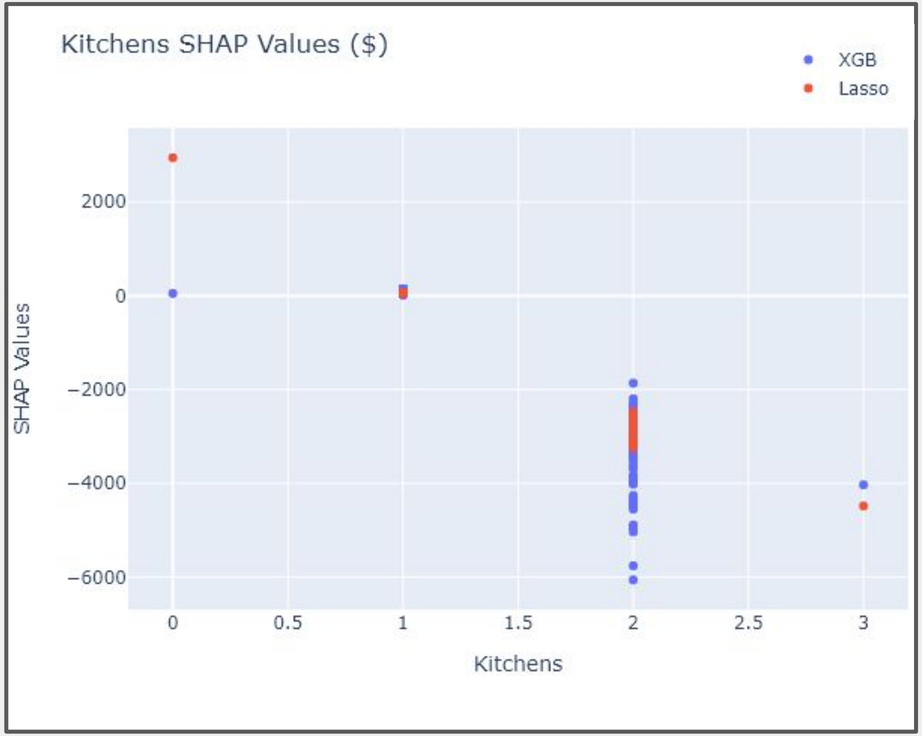


South West of Iowa State University

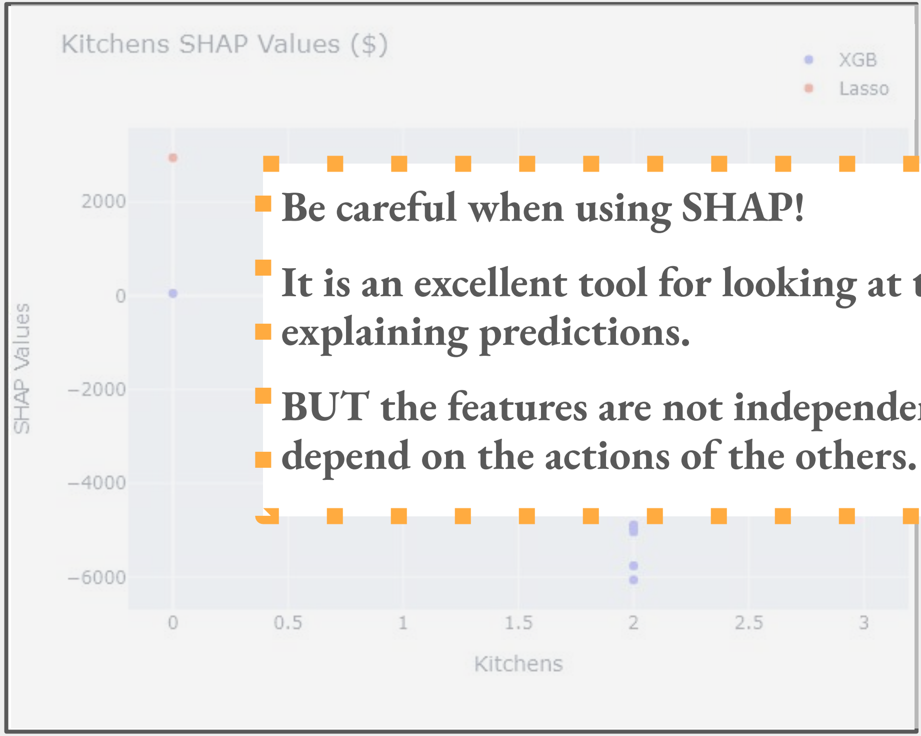


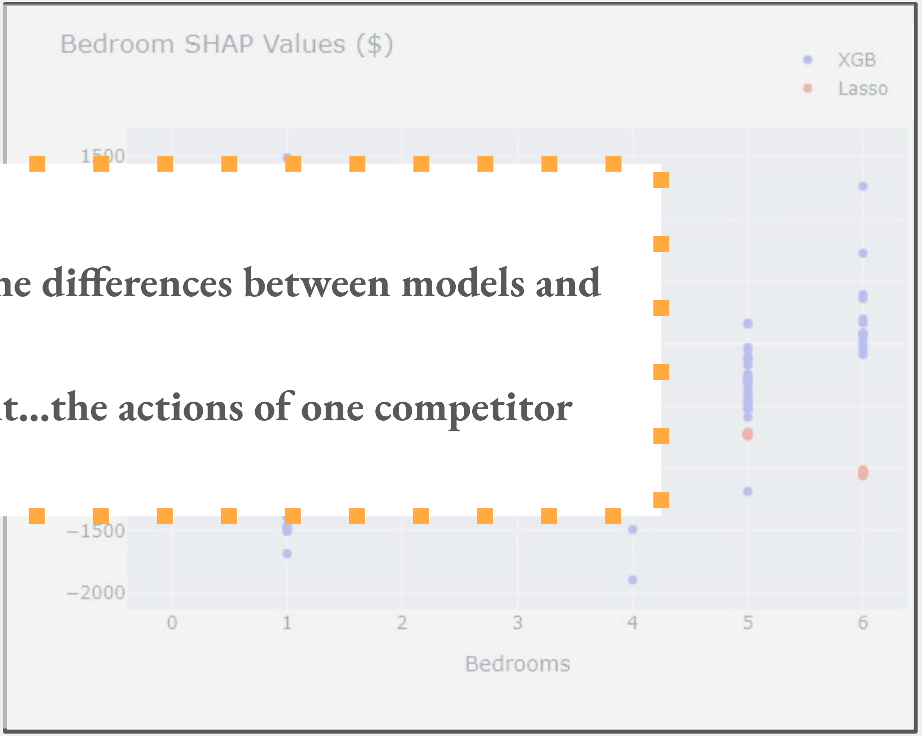
8

SHAP - Kitchens and Bedrooms?



SHAP - Kitchens and Bedrooms?





Be careful when using SHAP!

It is an excellent tool for looking at the differences between models and explaining predictions.

BUT the features are not independent…the actions of one competitor depend on the actions of the others.

Conclusions

What makes a valuable home in Ames, IA?

- ➔ Large homes with a large basement and garage.
- ➔ High home quality and condition.
- ➔ Newer homes, particularly those built after 1980.
- ➔ Certain neighborhoods increase home value.
- ➔ A brick exterior increases home value.
- ➔ Certain features provide diminishing return after a certain point
  - ♦ Lot area (20000 square feet)
  - ♦ Garage Cars (3 cars)
    (Big garage is always good)
  - ♦ Fireplaces (2)
  - ♦ Kitchens (only 1 above grade)
  - ♦ Bathrooms (3.5)

Future Work

More work with feature engineering

Optuna

- ➔ More hyperparameters: Input nulls/variables to drop
- ➔ Run through Optuna for each of the 8 dataframes to see how the ideal hyperparameters vary for each algorithm

SHAP

- ➔ Look at other tree models and see how they compare to XGBoost

Catboost/LightGBM

Incorporate geographic data

Acknowledgment

- Thank you Vinod!
- Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project by Dean De Cock
- Exploratory Data Analysis of Housing in Ames, Iowa by Lee Clemer
- Using optuna with sklearn the right way — Part 1 by Walter Sperat
- Using optuna with sklearn the right way — Part 2 by Walter Sperat
- Interpretable Machine Learning: A Guide For Making Black Box Models Explainable by Christoph Molnar
- Tidy modeling with R by Max Kuhn AND Julia Silge