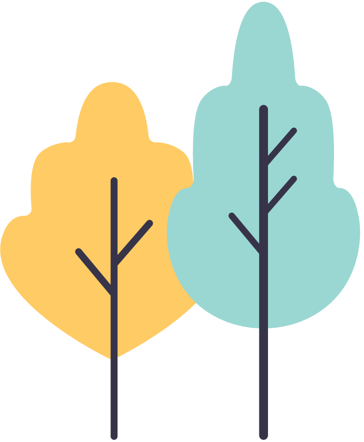
Predicting Housing Prices in Ames, Iowa

Kristin Teves
Van Vu

ABOUT THE DATASET

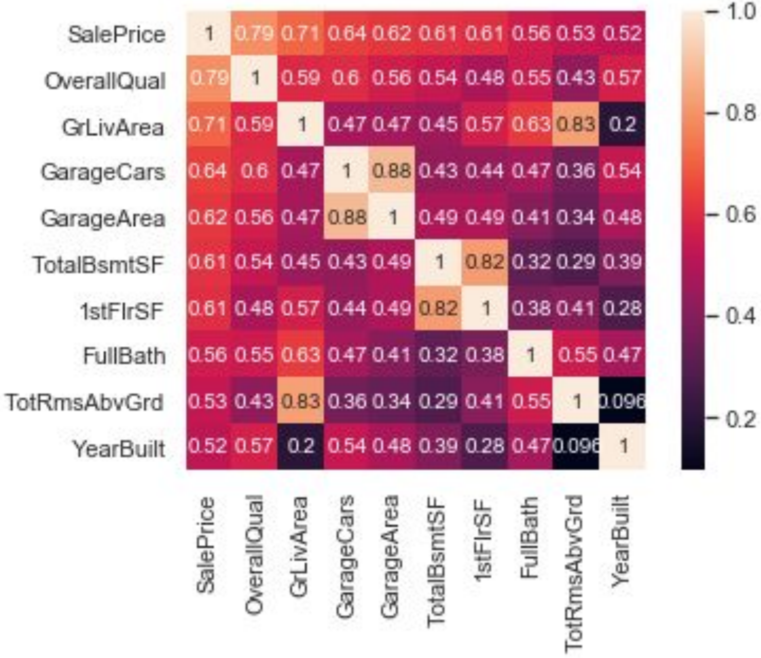
- Data set contains characteristics related to the residential homes in **Ames, Iowa** to describe sale prices between 2006 to 2010
- The **train** and **test** data sets were **combined** to include over 2900 observations and 81 features of nominal, ordinal, discreet, and continuous data type to assess home values

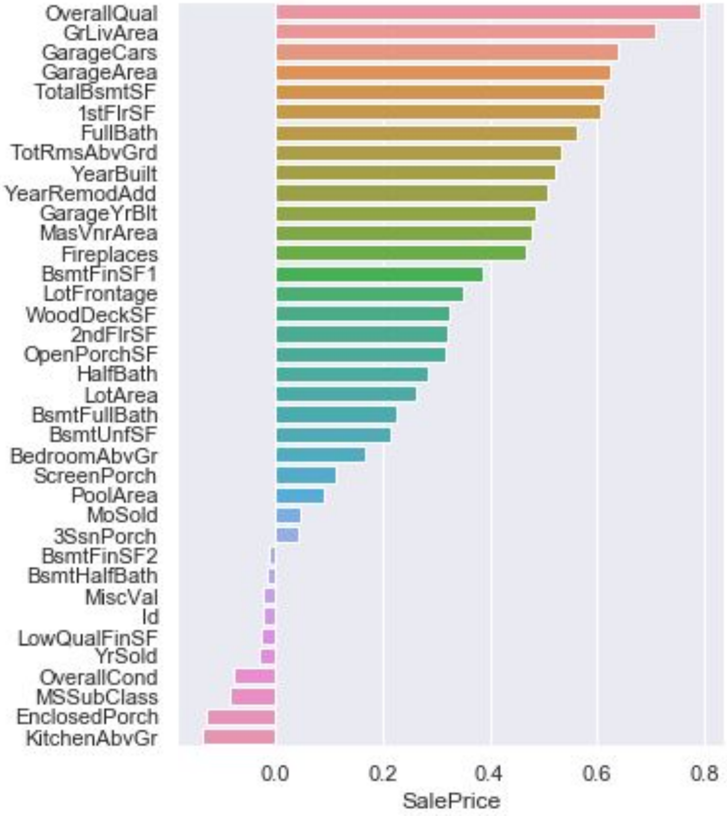


PROJECT GOALS

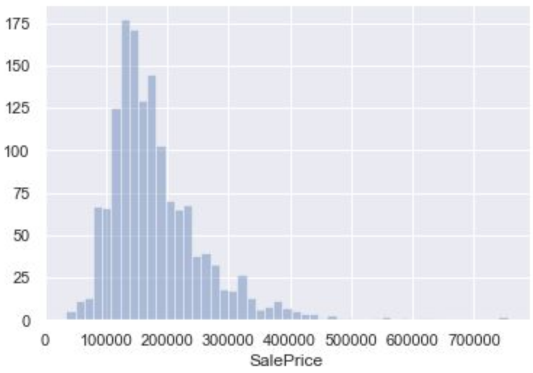
- Explore varying **regression models** and identify which model best predicts housing prices
    - Model is scored using R^2
    - Identify and engineer new features that help predict home values

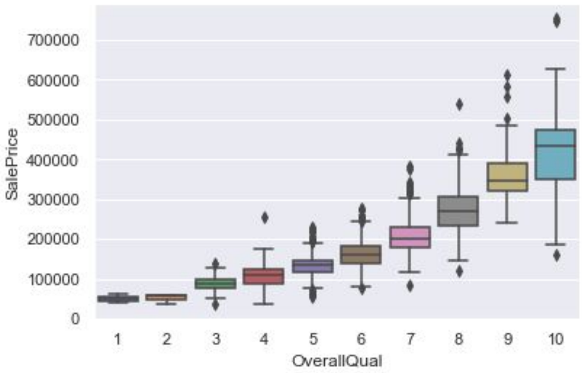
EXPLORING THE DATA

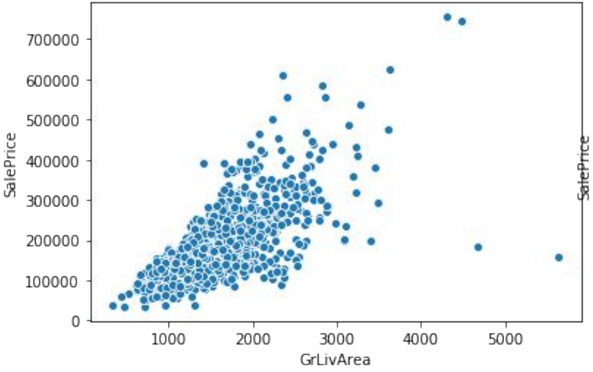


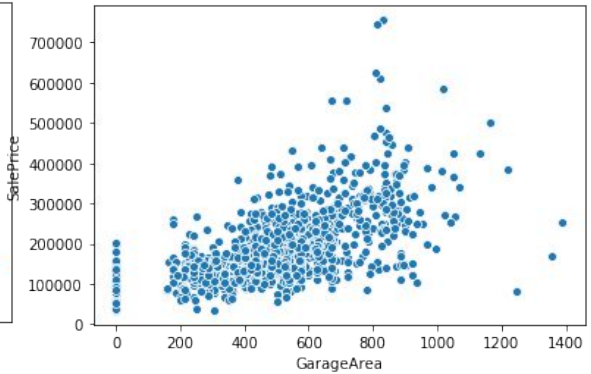


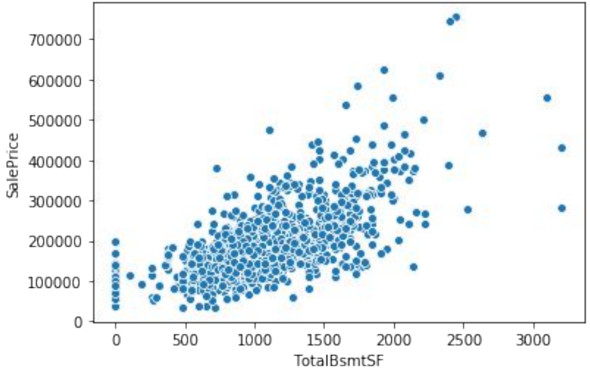
Exploring the Data











Pre-processing and Machine Learning

01
Missing Data

Identify features with missing data and decide what to do

02
Imputing Data

Decide what values to impute for missing data

03
Feature Engineering

Transform data into new features

04
Feature Selection

F Test

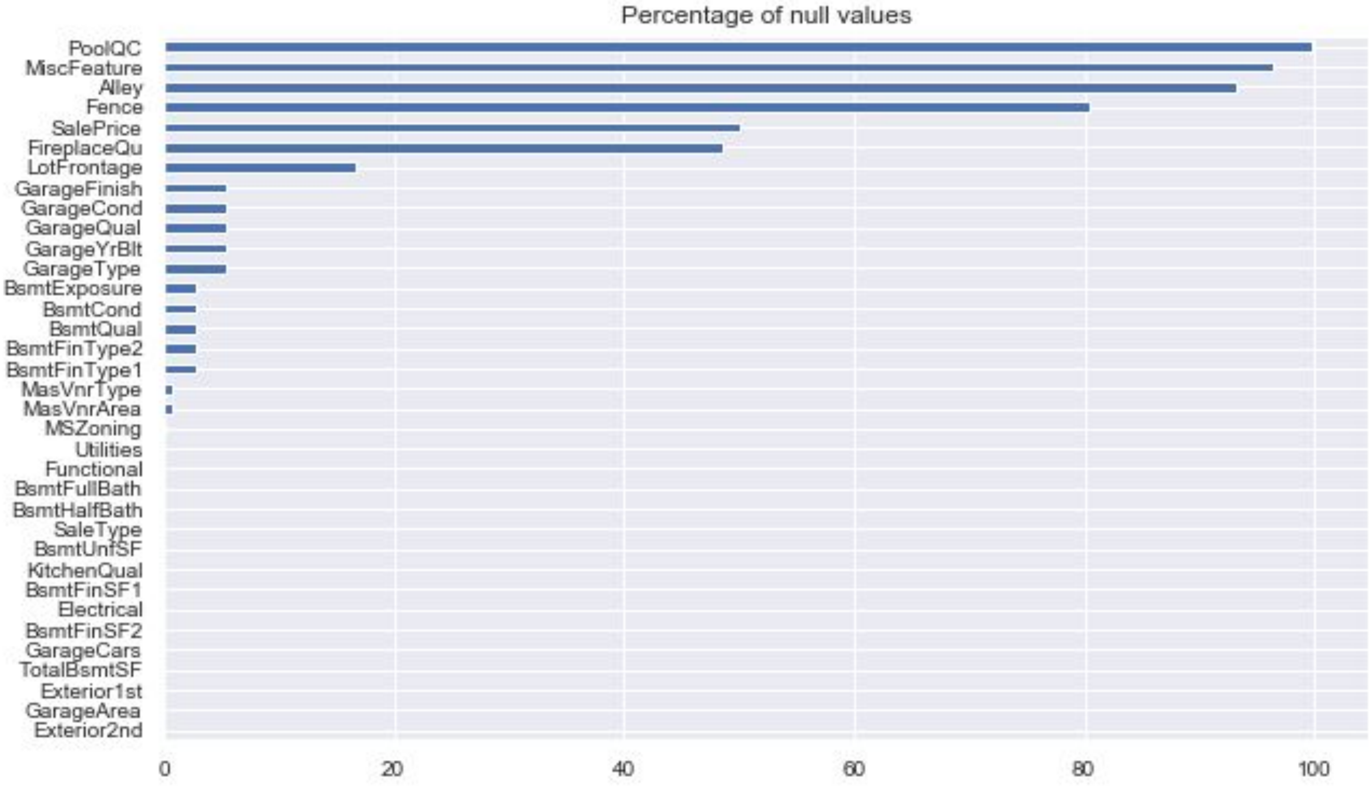
05
MLR, Ridge, and Lasso

Regression Models

06
Random Forest and Gradient Boosting

Advanced Regression Models

MISSING DATA



DELETING MISSING AND IRRELEVANT DATA

- Delete Id, which does not affect our price predictions
- Delete features with more than 80% missing values: PoolQC, MiscFeature (and its counterpart, MiscVal), Alley, Fence

IMPUTE MISSING DATA

Impute some categorical features with "**None**"

- GarageQual, GarageYrBlt, GarageFinish, GarageCond, GarageType, BsmtCond, BsmtExposure, BsmtQual, BsmtFinType2, BsmtFinType1, FireplaceQu, MasVnrType
- Advanced Model - **GarageCars, GarageArea, MasVnrArea

Impute some numerical features with **Zeros**

- BsmtFullBath, BsmtHalfBath, BsmtUnfSF, BsmtFinSF1, BsmtFinSF2, TotalBsmtSF
- Simple Model - **GarageCars, GarageArea, MasVnrArea

Impute some others with **mode**

- LotFrontage, MSZoning, Utilities, Electrical, KitchenQual, SaleType, Functional, Exterior1st, Exterior2nd

** Differences in dataset across models

FEATURE ENGINEERING

Numerical features were **combined** to form one new feature, and dropped afterwards. Dropped 12 of these features.

- TotalBaths = FullBath + (HalfBath * 0.5) + BsmtFullBath + (BsmtHalfBath * 0.5)
- PorchSF = WoodDeckSF + OpenPorchSF + EnclosedPorch + 3SsnPorch + ScreenPorch
- TotalSF = TotalBsmtSF + 1stFlrSF + 2ndFlrSF

Convert **MSSubClass** to **string** type since the numerical values identify the type of dwelling involved in sale, not ordinal numeric value

FEATURE ENGINEERING

Convert **quality** and **conditions** string values to **ordinal numerical** values

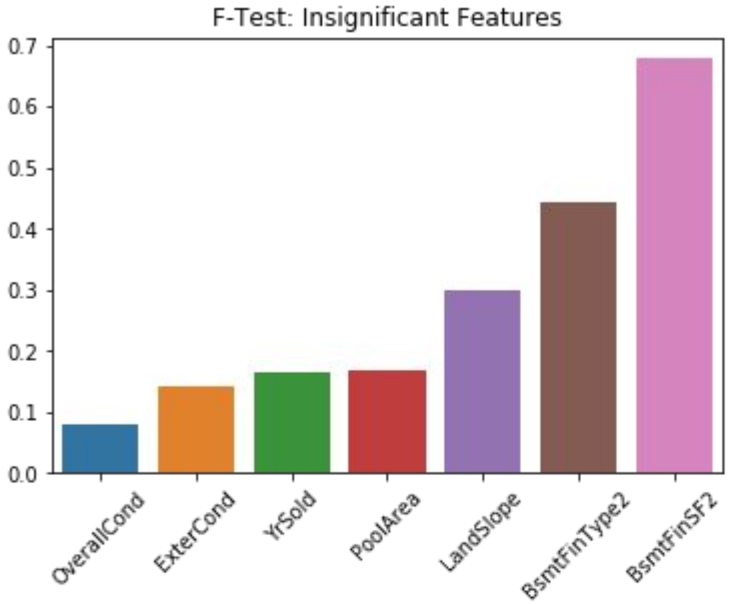
- ExterQual, ExterCond, BsmtQual, BsmtCond, HeatingQC, KitchenQual, GarageQual, GarageCond
    - {None: 0, Po: 1, Fa: 2, TA: 3, Gd: 4, Ex: 5}

Convert **other categorical** variables to **ordinal numerical** values

- LotShape - {IR3: 1, IRF2: 2, IRF1: 3, Reg: 4}
- BsmtExposure - {None: 0, No: 1, Mn: 2, Av: 3, Gd: 4}
- BsmtFinType1 and BsmtFinType2 - {None: 0, Unf: 1, LwQ: 2, Rec: 3, BLQ: 4, ALQ: 5, GLQ: 6}
- Functional - {None: 0, Sal: 1, Sev: 2, Maj2: 3, Maj1: 4, Mod: 5, Min2: 6, Min1: 7, Typ: 8}
- GarageFinish - {None: 0, Unf: 1, RFn: 2, Fin: 3}
- PavedDrive - {N: 0, P: 1, Y: 2}
- CentralAir - {N: 0, Y: 1}
- LandSlope - {Gtl: 1, Mod: 2, Sev: 3}

FEATURE SELECTION

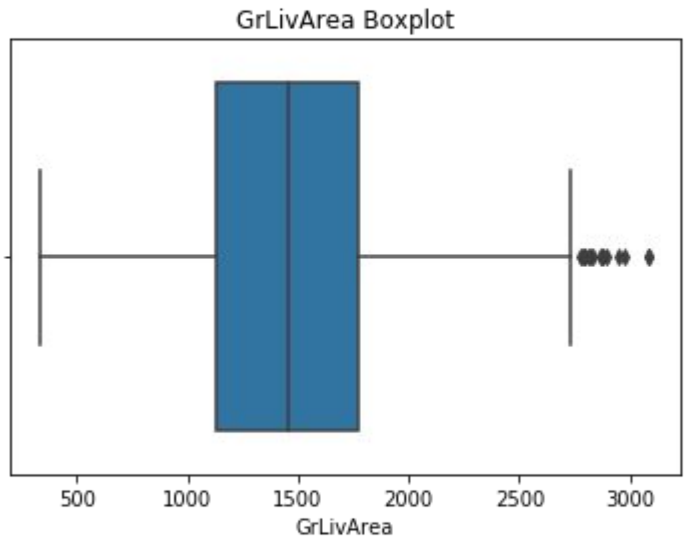
Used F-Test regressor to determine which coefficients are statistically significant to improve the fit of the model



OUTLIERS

Drop GrLivArea outliers, accepting z-score less than or equal to 3

```
TotalSF          0.000000e+00
OverallQual      0.000000e+00
GrLivArea        3.441763e-222
ExterQual        6.499902e-202
KitchenQual      2.037484e-187
TotalBaths       3.273788e-185
```

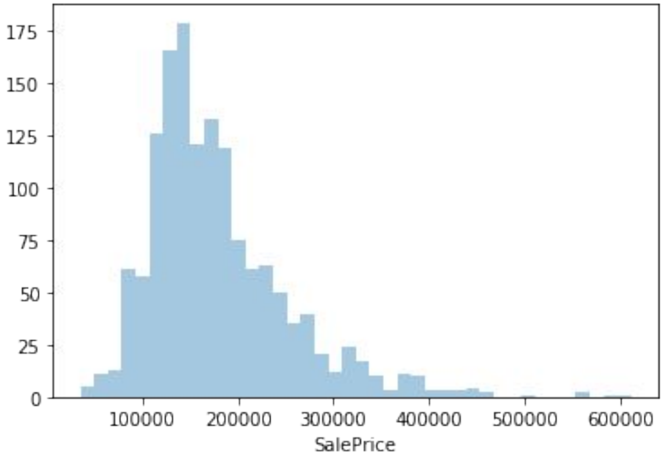


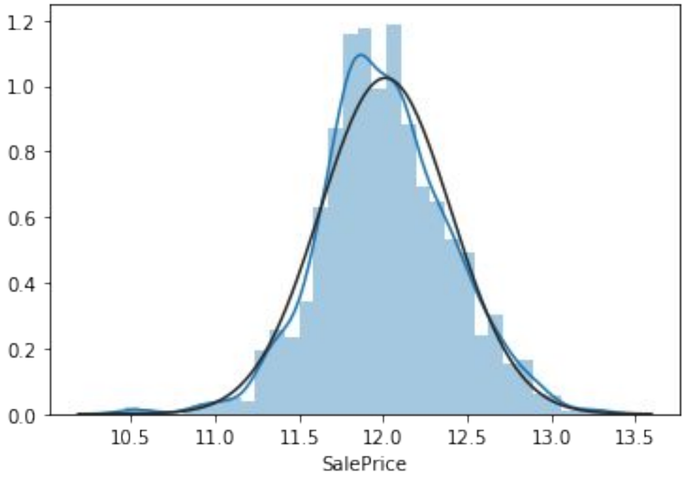
MODEL SELECTION

- Multiple Linear Regression
- Ridge, Lasso, Elastic Net
- Random Forest
- Gradient Boosting

LINEAR MODELS

- Made SalePrice normally distributed with log transformation to improve model fit





LINEAR MODELS

- Lasso feature selection resulted in 20 most significant features
- ElasticNet with alpha 0.001 and rho 0.6 performed the best of all the linear models

OverallQual	7.609837e-02
CentralAir	4.126533e-02
KitchenQual	3.688449e-02
MSZoning_RM	3.625271e-02
TotalBaths	3.294864e-02
Fireplaces	1.862262e-02
GarageFinish	1.680169e-02
GarageCars	1.271793e-02
ExterQual	9.478927e-03
FireplaceQu_None	2.377587e-03
YearRemodAdd	1.164068e-03
HeatingQC	9.787742e-04
GarageType_Attchd	5.477889e-04
YearBuilt	3.871165e-04
TotalSF	1.592307e-04
GarageArea	1.317889e-04
GrLivArea	3.190666e-05
BsmtFinSF1	1.864635e-05
PorchSF	1.284876e-05
LotArea	5.248903e-07

RANDOM FOREST

Random Forest Regressor

```
The training error is: 0.98019
The test     error is: 0.88428
```

GridSearchCV

```
grid_search_forest.best_params_
```

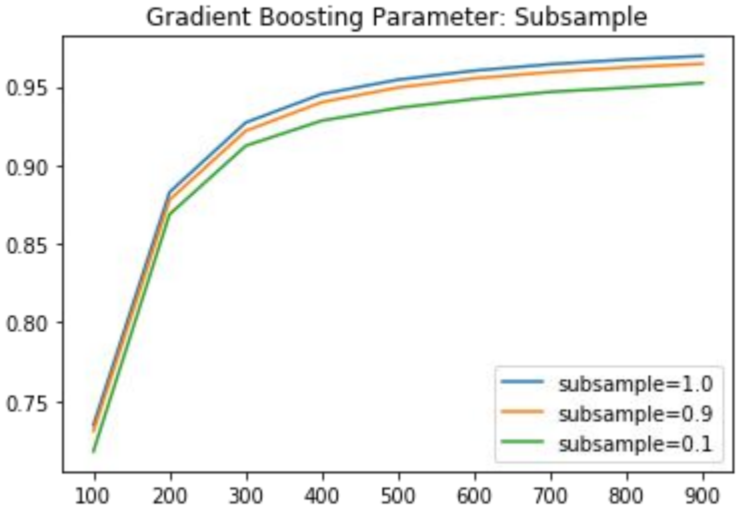
```
{'max_depth': 11, 'n_estimators': 550}
```

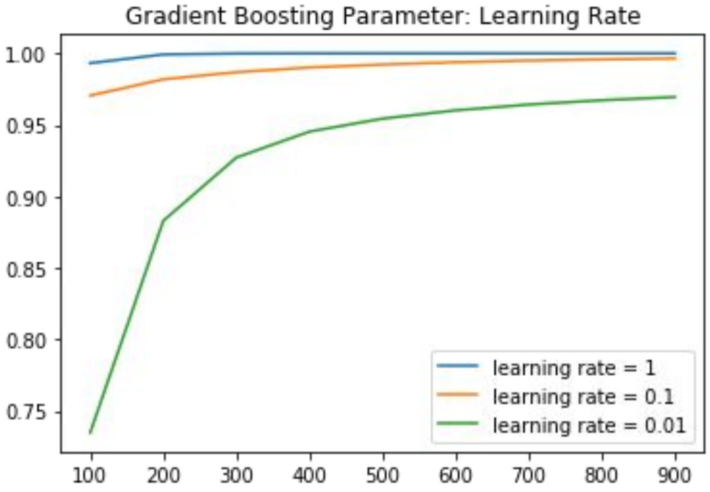
```
The training score is: 0.98313
The test     score is: 0.90266
```

GRADIENT BOOSTING

```
gbm.set_params(subsample = 0.9, n_estimators = 500)
```

```
GradientBoostingRegressor(alpha=0.9, criterion='friedman_mse', init=None,
                          learning_rate=0.1, loss='ls', max_depth=3,
                          max_features=None, max_leaf_nodes=None,
                          min_impurity_decrease=0.0, min_impurity_split=None,
                          min_samples_leaf=1, min_samples_split=2,
                          min_weight_fraction_leaf=0.0, n_estimators=500,
                          n_iter_no_change=None, presort='auto',
                          random_state=None, subsample=0.9, tol=0.0001,
                          validation_fraction=0.1, verbose=0, warm_start=False)
```





RESULTS

	TRAIN R SQUARED	TEST R SQUARED	KAGGLE SCORE
MLR	0.948	0.866	0.228
ELASTIC NET	0.92	0.92	0.141
RANDOM FOREST	0.984	0.902	0.159
GRADIENT BOOSTING	0.993	0.921	0.138

CONCLUSIONS

- Feature engineering played a key role in improving the accuracy score for each model.
  - Imputing data as “None” vs Zero
- Simpler regression models, like MLR, contained lower accuracy, or R^2, scores compared to Gradient Boosting.
- Advanced regression model, **Gradient Boosting Regression**, yielded the best R^2 and Kaggle Score. Therefore, Gradient Boosting predicted SalePrice best.

FUTURE WORK

- Minimize overfitting
  - Cross Validation
  - Tune parameters
- Improve feature selection - fit model with select features

THANKS

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik**.

Please keep this slide for attribution.