Group 1: Housing Prices

Klement Florian, Minaeva Anastasiia, Pollek Patrick, Trush Maria

The Problem

Annually up to 1 million houses are sold in the US. How can Data Science help in buying the best house at the right price?

Main Goals

#1 **Predict price of a house**

- Build a user-friendly model based on readily available parameters
- Minimize the error and potential losses for the buyer/seller

#2 **Determine quality category of a house**

- Does a house fall into a certain desired category?
- Predict quality of a house

The Dataset

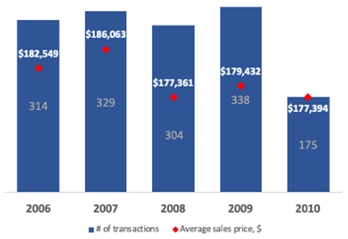
1,460 home sales transactions in the city of Ames, Iowa from **2006** to **2010**

81 attributes, including **43 categorical attributes**, such as:

- House type, street access, neighborhood, materials, basement quality, garage type, porch, sales type, electrical system

And **38 numerical attributes**, such as:

- Lot area, 1st floor area, # of bathrooms, # of bedrooms, year built

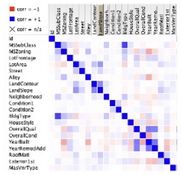


Houses Sales Prices SummaryStatistics	
Mean price, $	$180,921
Max price, $	$755,000
Min price, $	$34,900
Standard deviation	$79,443

Data Understanding & Preparation

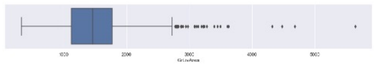
Correlation

- Not highly correlated; only 4 pairs of attributes with correlation over 80%



Correlation to Sales Price	
Overall Quality	0.79
Gr. Living Area	0.71
Garage Cars	0.64
Basement Area	0.61

Outlier Treatment



- Boxplots identified too many outliers
- Opt for Automatic Outlier detection (Isolation forest)

Attributes Reduction

- Attributes with high correlation
- Attributes where >80% of values are the same (reduced from 81 to 44 attributes)

Normalization & Scaling

- Min-Max normalization

Modeling

#1 House Price Prediction

5 Models Used

Model	Mean Sq Error (Test)[1]
ElasticNet	0.0139
RandomForestRegressor	0.0214
SGDRegressor	0.0146
KNeighborsRegressor	0.0341
MLPRegressor (neural network)	0.0159

Final Model

- Predict the house price as a mean of 5 models
- Resulting mean error **US$20,608**

[1] MSE for scaled results

#2 Quality Ranking Prediction

Unsupervised & Supervised Learning

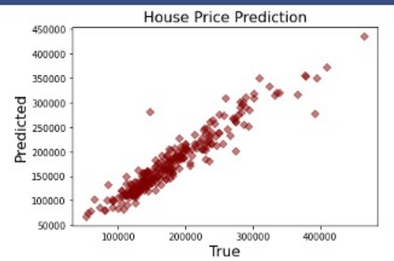
- **Unsupervised learning**: clustering. DB Scan did not produce meaningful results
- **Supervised Learning:** 4 models applied, including (i) Logistic Regression, (ii) Random Forest, (iii) KNeighbors, and (iv) SVC. Accuracy ranges from 51% to 58%.

Final Model

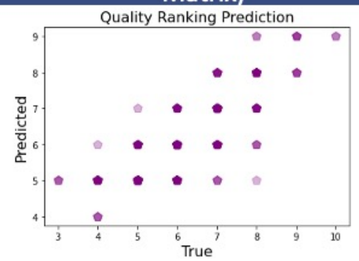
- Predict the house quality ranking (1-10) using Random Forest with **accuracy of 59%**

Results & Conclusions

Predicted Prices



Predicted Quality Category (Confusion Matrix)



- **House Price Prediction**: Final model produces relevant results with **mean error of US$20,608,** compared to **mean house price of US$180,921**. Can be used for housing market analysis and for buy/sell decision, especially to identify under/overpriced houses.
- **Quality Ranking Prediction**: Although overall accuracy of 59% is relatively modest, the model predicts a quality category within **+/-1 interval in 96%** of cases, providing an important metric for marketing and selecting a house.

Workflow