# Realty Mogul: Real Estate Price Prediction with Regression and Classification

Hujia Yu, Jiafu Wu, [hujiay, jiafuwu]@stanford.edu

## Motivation

The Ames Assessoris Office released information on its sold houses from 2006 to 2010. Housing prices are an important reflection of the economy, and houses' price ranges are of great interest for both buyers and sellers. In this project, sale prices will be predicted based on a variety features of residential houses both as a continuous response variable and multinary response variables, with classifications determined by the following price ranges:

[0, 100K), [100K, 150K), [150K, 200K), [200K, 250K), [250K, 300K), [300K, 350K), [350K, inf)

## Data and Features

Dataset: residential houses in Ames, Iowa sold in 2006 - 2010

- 79 house features
- 1460 houses with sold prices

Preprocess the data:

- Turn categorical data into separated indicator data.
- Fill in null value as 0 indicator value
- Randomly select training and testing examples among 1460 examples.
- Set aside sold prices in testing examples as ground truth
- Sale Price is log transformed to have a normalized distribution during regression analysis

Final dataset

- 288 house features
- 1000 training examples
- 460 testing examples.

## Models

### Classification

Naive Bayes (Gaussian/Multinomial)

$$\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),$$

Multinomial Logistic Regression

$$\min_{w,c} \|w\|_1 + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1).$$

SVM Classification (Linear/ Gaussian)

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \zeta_i$$

$$\text{subject to } y_i(w^T \phi(x_i) + b) \geq 1 - \zeta_i,$$

$$\zeta_i \geq 0, i = 1, ..., n$$

Random Forest Classification

Constructing a multitude of decision trees at the training time and output the decision of the class at test

### Dimensionality Reduction

Principal Component Analysis

$$\hat{u}_{(1)} = arg \max_{||u||=1} u^T X^T X u$$

$$\hat{X}_{(k)} = X - \sum_{s=1}^{k-1} X \hat{u}_{(s)} \hat{u}_{(s)}^T$$

$$\hat{u}_{(k)} = arg \max_{||u||=1} u^T \hat{X}_{(k)}^T \hat{X}_{(k)} u$$

### Regression

Ridge Regression

$$\min_w ||Xw - y||_2{}^2 + \alpha ||w||_2{}^2$$

Lasso Regression

$$\min_w \frac{1}{2n_{samples}} ||Xw - y||_2^2 + \alpha ||w||_1$$

SVM Regression

Similar to SVM Classification

Random Forest Regression

Similar to Random Forest Classification

## Results

| Classification Model | Classification Error Rate | Classification Model w/ PCA | Classification Error Rate | Regression Model | Regression RMSE |
|---|---|---|---|---|---|
| Gaussian Naive Bayes | 0.7913 | PCA + Gaussian Naive Bayes | 0.5022 | Linear Regression | 0.5501 |
| Multinomial Naive bayes | 0.4891 | - | - | Lasso | 0.4954 |
| Multinomial Logistic Regression | 0.500 | Multinomial Logistic Regression | 0.4413 | Ridge | 0.5448 |
| SVC linear kernel | 0.3260 | SVC linear kernel | 0.3087 | SVR (linear kernel) | 5522 |
| SVC Gaussian kernel | 0.5891 | SVC Gaussian kernel | 0.5891 | SVR (Gaussian kernel) | 0.5016 |
| Random Forest Classification | 0.3348 | Random Forest Classification | 0.4326 | Random Forest Regression | 0.5394 |

## Discussion

- **Classification:** We treated Gaussian Naive Bayes as baseline and it performed poorly with 0.79 error rate. The best models for these classification problem include SVC with linear kernel and random forest. One possible cause of the error might be that there are too many features (288) and it leads to overfit. We use PCA for dimensionality reduction and it indeed improved the performance of the models.
- **Regression:** We treated linear regression with all covariates as baseline, and it generated RMSE of 0.5501. Overall, most of the regression models gave better results than our baseline model, except SVR with linear kernel, which is not innately suitable for fitting linear regression data set like this. Linear regression with Lasso turned out to perform the best due to its feature reduction function. According to our model, the year that the house was built turned out to have the greatest statistical significance upon predicting the sale price of a house.

## Future

- The number of covariates existent in our dataset is abundant, but feature selection helped constrain the complexity of our models in this setting.
- With around 0.3087 error rate, our SVC with linear kernel model could be used for price range predictions for future houses in Ames, Iowa.