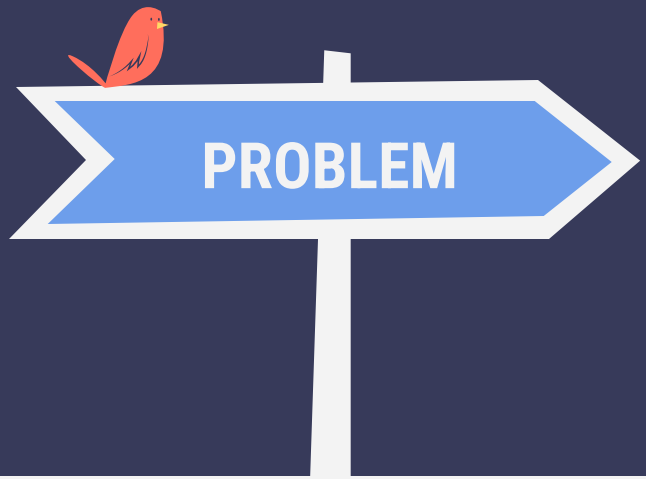
PREDICTING REAL ESTATE PRICES

CS 377 Machine Learning

Kristina K., Serenity S., Justin K., Omari B.

As the housing market increases in prices, homebuyers have a difficult time finding their dream home.



Using a homebuyers preferences, we use OLS (Ordinary Least Squares) Regression to help find the best home for them.

PROJECT EXECUTION

01.
Classify the question/problem statement

02.
Cleaning the dataset

03.
Choose our regression model

04.
Compare with other models

05.
Test the data

06.
Accurately provide person with their dream home

- **SCRUBBED THE DIRTY DATA!!!** -

```
houses = houses.drop(['ID'], axis=1)
```

```
current_year = 2024
houses['Age'] = current_year - houses['Year_Built']
```

WHY REGRESSION?

1

Three types of Machine Learning:
1. Reinforcement: trial/error
2. Unsupervised: finding patterns in unlabeled data
3. Supervised: predicting outcomes using labeled data

2

We know our input and output variables so Supervised learning it is!

3

Now the question is: Classification or Regression?

4

The Answer is: Regression, since we're predicting a continuous variable and not classes

REGRESSIONS WE RAN

OLS

- **Ordinary Least Squares:** to find the best “fit” line
- To minimize the sum of the square differences (errors)

94.94%

LASSO

- **Lasso Regression:** linear regression model that adds a penalty to prevent overfitting

94.93%

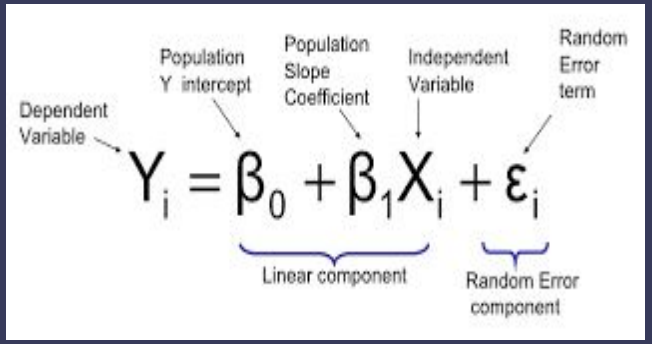
RIDGE

- **Ridge Regression:** linear regression model primarily used for data with multicollinearity

94.91%

MATH FOR OLS REGRESSION

Ordinary Least Squares



ANALOGY: Finding the straightest path through a city where houses represent data points. Goal is to minimize the total squared distance (errors) between the houses (actual values) and your path (predicted values) ensuring route fits data as closely as possible.

- **Goal:** to minimize prediction error
  - the residual sums of squares (RSS); measure of how far the predicted values from a model are from the actual data points
- Uses unbiased estimators (no data assumptions)
- Trade-off: High variance -> predictions change a lot based off of slight changes in the data as it's a more complex model
- Retains all features, even irrelevant ones

MATH FOR LASSO REGRESSION

Least Absolute Shrinkage & Selection Operator

$$\sum_{i=1}^{n}(y_i - \sum_{j} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

ANALOGY: Packing your suitcase for a trip, but there's a weight limit (λ). You want to bring your favorite items (fit the data well) but if something isn't very useful (low importance), you leave it behind to avoid exceeding the weight limit (penalty).

- **Goal**: Minimize error, simplify model
- Error = Sum of Squared Residuals + λ · (Sum of Absolute Coefficients)
- Sum of Squared Residuals → ensures model fits the data by penalizing outliers
- λ · (Sum of Absolute Coefficients) -> penalizes large coefficients by shrinking many to **exactly** zero
  - If λ = 0 then it's identical to OLS
  - If λ is large then it dominates, shrinking the model
- Bias increases and variance decreases as penalty (λ) gets larger sacrificing accuracy on training data to improve generalization on unseen data and making it less sensitive to noise
- Less prone to overfitting

MATH FOR RIDGE REGRESSION

RIDGE

$$RSS_{L2} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^{P} B_j^2$$

ANALOGY: Packing a suitcase with a weight limit. You can bring everything (all features), but you shrink/minimize the bulk (coefficients) of each item to fit within the limit. This prevents overpacking (overfitting) while keeping everything useful.

- **Goal**: Minimize error
- Error = Sum of Squared Residuals + λ · (Sum of Squared Coefficients)
- Sum of Squared Residuals → ensures model fits the data by penalizing outliers
- λ · (Sum of Squared Coefficients) → penalizes the large coefficients
  - If λ = 0 then it's identical to OLS
- Shrinks all coefficients but targets large ones and keeps them nonzero
  - Less prone to overfitting and reduce noise sensitivity
- Handles multicollinearity -> when 2 or more independent variables are highly correlated with each other
  - OLS struggles alone because it amplifies noise and produces overfitting

ATTRIBUTES OF REGRESSION

.Number of bedrooms

.Number of bathrooms

.Desired square feet

.Number of floors

.Garden

.Pool

.Age of house

.Garage size

```
user_bedrooms = float(input('How many bedrooms do you want? '))
user_bathrooms = float(input('How many bathrooms do you want? '))
user_square_feet = float(input('How large do you want your home to be in sq. ft? '))
user_floors = float(input('How many floors do you want? '))
user_garden = float(input('Do you want a garden? '))
user_pool = float(input('Do you want a pool? '))
user_age = float(input('How old do you want the house to be? '))
user_garage = float(input('How large do you want the garage to be? '))
```

FORMULA FOR OLS

Price = -2809099.35 + ***(987.42)Square_feet*** + ***(49553.32)Num_bedrooms*** + ***(31084.16)Num_bathrooms*** + ***(17691.23)Num_floors*** + ***(-1502.79)age*** + ***(29464.54)Has_garden*** + ***(39902.11)Has_pool*** + ***(1135.51)Garage_size***