

ECOM90025 Assignment 5

Team Member:

Keyuan Zhang 1378707

Sissie Zhang 1423417

Zheyuan Li 1342654

Zhicheng Liu 1173419

Guozheng Tian 1393719

1. Abstracts

This project emphasizes advanced data preprocessing techniques. By employing interaction terms and principal component analysis, we systematically reduce the dimensionality of the dataset and enhance model accuracy. The optimal number of components is determined via the Elbow method, followed by a regression incorporating regularization techniques.

2. Interaction Terms Creation

Prior to the implementation of principal component analysis, we aim to enhance the model's capability to elucidate nonlinear relationships and optimize its efficiency. To this end, the project identify the top five variables exhibiting the highest correlation with sales prices and subsequently generate two-way interaction terms for them. These interaction terms are integrated into the dataset without replacement.

Feature	Correlation
OverallQual	0.8172
TotalArea	0.8071
GrLivArea	0.7303
GarageCars	0.6806
GarageArea	0.6509

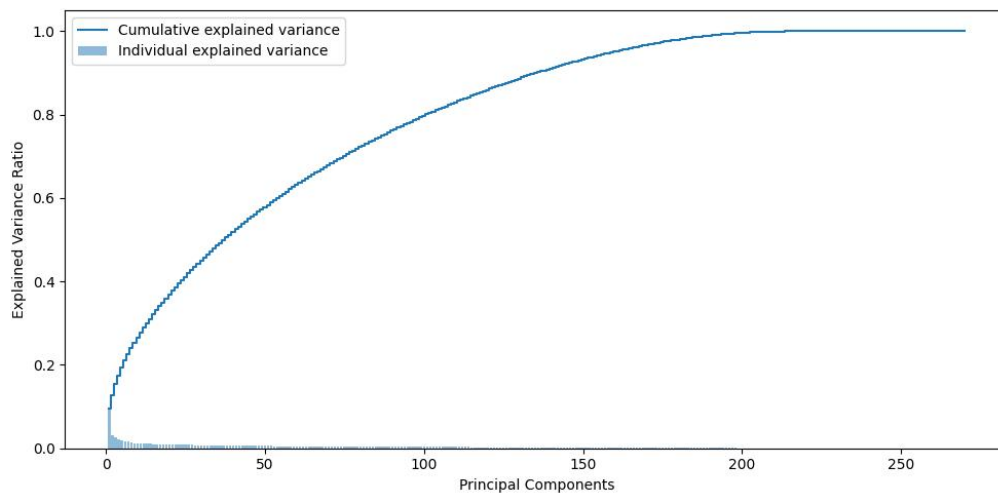
Table 1: Five Features with the Highest Correlation with Sales Price

3. PCA Processing

PCA is pivotal in data preprocessing. It converts high-dimensional datasets to a simpler low-dimensional form, reducing model complexity. It also filters out less important components, preserving key dataset features with minimal information loss.

To address potential biases introduced by varied scales, all data points are standardized to achieve uniform variance across the dataset. During the PCA, all components are initially retained to encompass the entire explained variance. This

ensures that the model captures the entirety of the dataset's information, providing a comprehensive foundation for subsequent component selection.

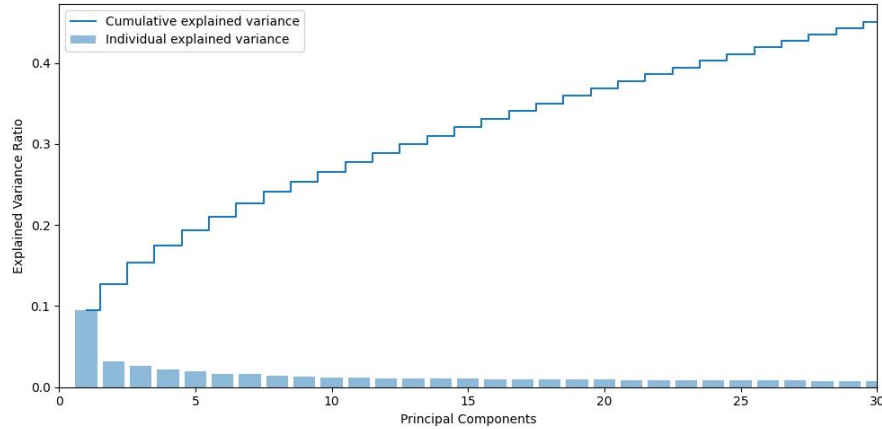


Plot 1: Explained Variance Distribution for all components

4. Components Selection with ELBOW

The ELBOW method is employed to guide the selection of the number of components. Components beyond the 'Elbow point' typically offer less explained variance, contributing minimally to the overall model. Consequently, this project identifies the third component as the 'Elbow point', leading us to retain the initial three components. Subsequent OLS regression between these components and housing prices reaffirmed a significant relationship between them.

In order to improve the interpretability and to include more information from original features, the project merged original features with the chosen components for regression. This approach is preferred over a direct substitution, ensuring a more comprehensive representation of the dataset in the regression.



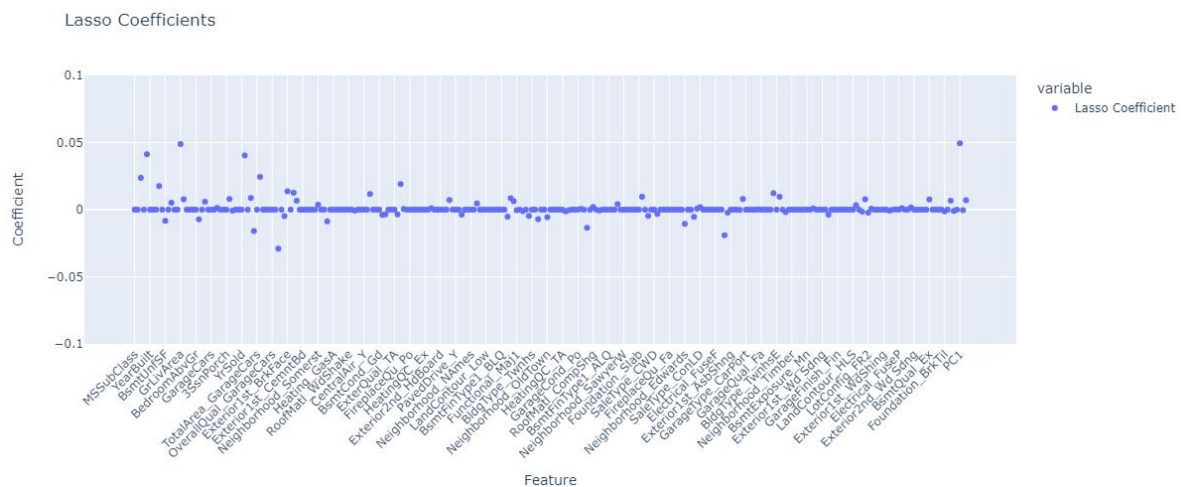
Plot 2: Explained Variance Distribution for first 30 components

	coef	std err	t	P> t	[0.025	0.975]
Intercept	12.0241	0.004	2811.37	0	12.016	12.032
PC1	0.0711	0.001	82.603	0	0.069	0.073
PC2	-0.0092	0.001	-6.291	0	-0.012	-0.006
PC3	0.0308	0.002	19.949	0	0.028	0.034

Table 1: OLS Regression between Selected Components and House Price

5. Regression with Lasso Regularization

To counteract multicollinearity and overfitting challenges, Lasso regression is employed, enhancing the model's precision. Upon regression with the PCA-processed data, the model exhibited great performance.

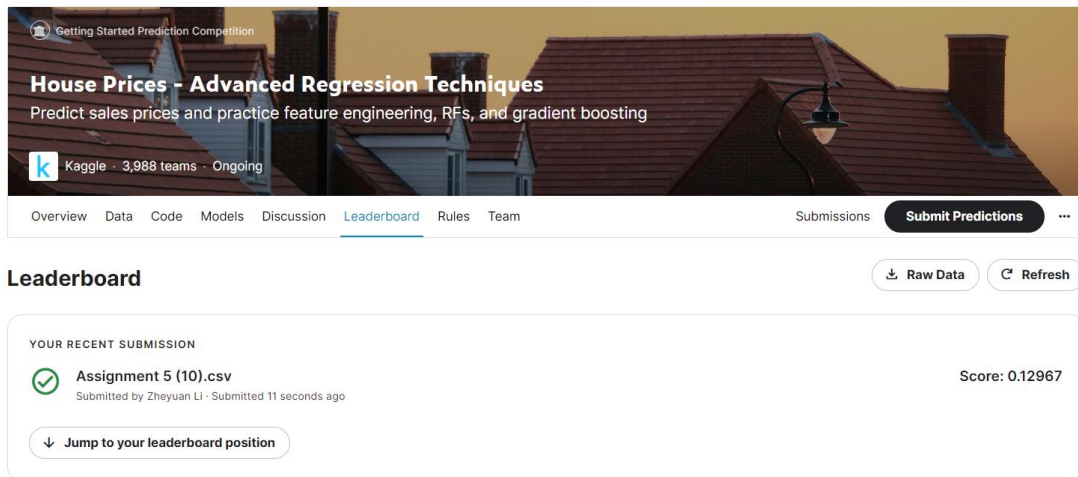


Plot 3: Lasso Coefficients Distribution

Python Code Link:

https://colab.research.google.com/drive/1zy1DZOVFCKnjs37VcwSZq8OTy-4_Rt6n?oid=114987379577149076166&usp=drive_link

Kaggle Score:



The image shows the Kaggle competition page for "House Prices - Advanced Regression Techniques". The header features a banner with a house roof and the text "Getting Started Prediction Competition", "House Prices - Advanced Regression Techniques", and "Predict sales prices and practice feature engineering, RFs, and gradient boosting". Below the banner, it says "Kaggle · 3,988 teams · Ongoing". The navigation bar includes "Overview", "Data", "Code", "Models", "Discussion", "Leaderboard" (which is highlighted), "Rules", and "Team". On the right, there are "Submissions" and a "Submit Predictions" button. Below the navigation bar, the "Leaderboard" section is visible, showing "YOUR RECENT SUBMISSION" with a green checkmark icon, the file name "Assignment 5 (10).csv", the submitter "Submitted by Zheyuan Li", the time "Submitted 11 seconds ago", and the score "Score: 0.12967". A button "Jump to your leaderboard position" is also present.