FEATURE EXTRACTION USING PCA AND FEATURE SELECTION USING LASSO

ON CAR DATASET

Prepared by:

Dhruvi Mehta (202318003)
Prachi Mehta (202318008)
Riya Dave (202318011)
Simran Dalvi (202318042)

INTRODUCTION

- In many applications, datasets often contain a large number of features, leading to challenges such as increased computational complexity, overfitting, and decreased model interpretability.
- The problem at hand is to efficiently extract relevant information from high-dimensional data while preserving the most important characteristics for accurate modeling.
- This necessitates the exploration of dimensionality reduction techniques such as feature extraction using PCA and feature selection using Lasso.
- The goal is to find an optimal balance between reducing the number of features and retaining the most informative ones to improve model performance and facilitate meaningful insights.

INTRODUCTION

- As a method for feature extraction, PCA reduces the size of a dataset without sacrificing important information.
- This kind of dataset compression makes PCA easier to use in later analyses and improves the visibility of underlying patterns that might have been hidden in the high-dimensional space at first.
- Lasso Regression is a feature selection technique that works in unitedly with PCA to help identify the most relevant variables for prediction.
- Lasso Regression encourages sparsity in the model by driving unnecessary feature coefficients to zero, which improves interpretability and may even improve predictive accuracy.
- Our goal is to improve fore cast accuracy and more easily understand the underlying causes of performance outcomes by combining the strengths of these two approaches.
- By means of a thorough compared study with traditional techniques, we aim to highlight the advantages and adaptability of our suggested methodology in various data analysis scenarios.

MATHEMATICAL FORMULATIONS AND INTUITION

- **1. Feature Extraction using Principal Component Analysis (PCA):**
- Centering the Data: Subtract the mean of each feature from the dataset to center the data:

$$X_{centered} = X - \mu$$

- Compute Covariance Matrix: Calculate the covariance matrix to understand the relationship between different features:

$$\Sigma = \frac{1}{n} X_{centered}^T X_{centered}$$

- Eigen Decomposition: Decompose the covariance matrix into its eigenvectors and eigenvalues to identify the principal components:

$$\Sigma V = V \Lambda$$

- Select Principal Components: Choose the principal components that capture the most variance in the data while reducing dimensionality.
- Project Data onto Principal Components: Transform the original data into a new space defined by the selected principal components:

$$X_{PCA} = X V_k$$

MATHEMATICAL FORMULATIONS AND INTUITION

- **2. Feature Selection using Lasso Regression:**
- Formulate Objective Function: Define the objective function for Lasso Regression, which includes a data fitting term and a regularization term:

$$\hat{\beta} = \arg\min_{\beta} \left(\frac{1}{2n} \|y - X_{PCA}\beta\|_2^2 + \lambda \|\beta\|_1 \right)$$

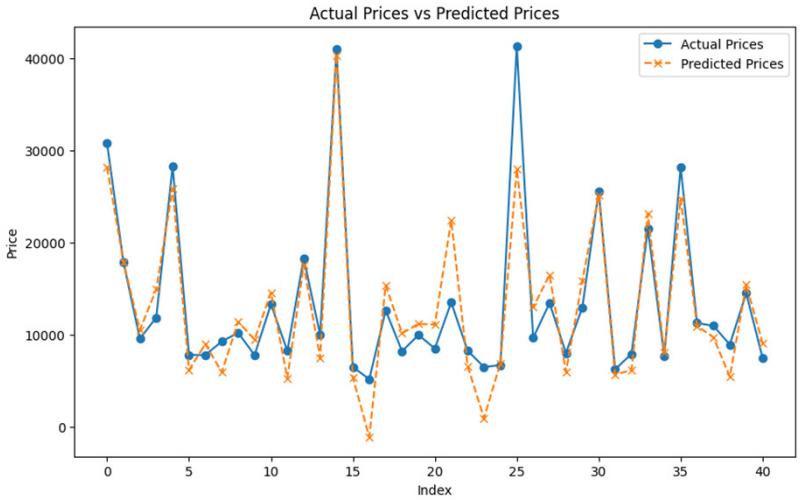
- Train Lasso Regression Model: Train the Lasso Regression model on the data to learn the optimal coefficients for each feature.
- Select Significant Features: Use the non-zero coefficients obtained from Lasso Regression to identify the most significant features.

MATHEMATICAL FORMULATIONS AND INTUITION

- **3. Model Training and Prediction:**
- Split Data: Split the dataset into training and testing sets to train and evaluate the predictive model.
- Train Prediction Model: Train a predictive model (e.g., linear regression, support vector machine) using the selected features as input.
- Predict on Test Data: Use the trained model to make predictions on the test dataset.

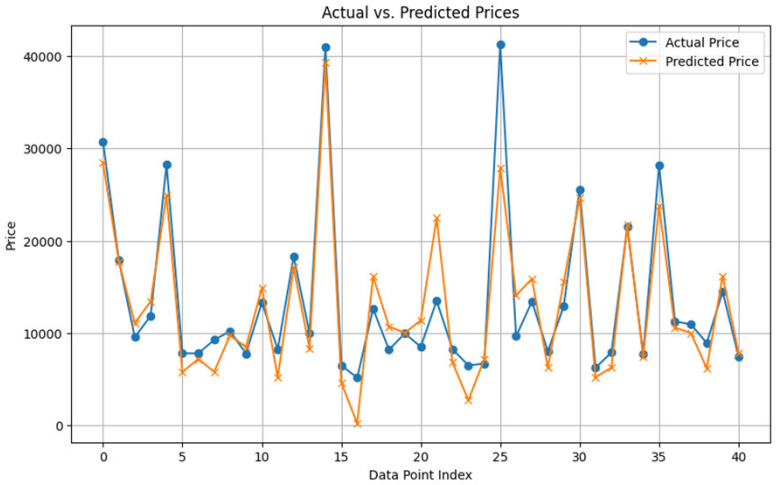
OBSERVATIONS

- Predicted price using Feature Extraction with PCA



The graph shows actual prices matching predicted prices closely, even with the use of PCA, validating the model's accuracy in forecasting.

- Predicted price using Feature Selection with Lasso Regression



Actual and predicted prices closely match, even with Lasso regression, affirming the model's accuracy in forecasting.

INSIGHTS

- Similar accuracy obtained from both PCA and Lasso regression suggests their effectiveness in capturing essential features influencing price dynamics.
- The consistency in accuracy between PCA and Lasso regression indicates that both methods extract or select relevant features sufficiently for price prediction.
- The close alignment between actual and predicted prices in the graph for both PCA and Lasso regression validates the accuracy of the models in forecasting prices.
- This alignment confirms that the chosen features, whether extracted through PCA or selected via Lasso regression, effectively capture the underlying patterns in the data related to price dynamics.
- The robustness and generalizability of the models from both PCA and Lasso regression enhance their utility in decision-making processes related to pricing strategies or market analysis.

CONCLUSION

- In conclusion:
- Both PCA and Lasso regression demonstrate similar accuracy in predicting prices, suggesting their effectiveness in capturing essential features.
- The close alignment between actual and predicted prices in the graphs for both PCA and Lasso regression confirms the reliability of the models in forecasting prices.
- This consistency in accuracy and alignment underscores the robustness of the models, regardless of whether features are extracted through PCA or selected via Lasso regression.
- These findings imply that either method can be employed with confidence for price prediction tasks, providing valuable insights for decision-making in pricing strategies and market analysis.

REFERENCES

- Dataset: https://github.com/prachii-15/Feature-selection-and-feature-extraction/blob/main/CarPrice_Assignment.csv
- Feature Extraction using PCA
- Feature Selection using Lasso Regression

THANK YOU…