DASC-5301-002-DATA SCIENCE
ASSIGNMENT-3

FEATURE SELECTION REPORT

Submitted by :-

ARHAAM DANIYAL SYED

ID:-1002169829

Masters in Data Science

The University of Texas at Arlington

THE DATA:

The dataset is titled "1985 Auto Imports Database." It contains information about automobiles, including various characteristics, insurance risk ratings, and normalized losses. The dataset consists of 205 instances (entries) and 26 attributes (features).

We have used our df2 which was already filtered in the previous homework which contains the features:

Wheel Base: The distance between the centers of the front and rear wheels on the same side of the vehicle.

Length:: The overall length of the vehicle.

Width:The width of the vehicle, measured from one side to the other.

Height:The vertical measurement of the vehicle from the ground to the highest point.

Curb Weight:The weight of the vehicle without any occupants or cargo but with all standard equipment and fuel.

Engine Size:The total volume of all cylinders in the engine.

Bore:The diameter of the cylinders in the engine.

Stroke:The distance the piston travels inside the cylinder.

Compression Ratio: The ratio of the maximum to minimum volume in the combustion chamber of an internal combustion engine.

Horsepower: A unit of power that measures the rate at which work is done by the engine.

Peak RPM (Revolutions Per Minute): The engine speed at which it produces the maximum power.

City MPG (Miles Per Gallon): The estimated fuel efficiency in miles per gallon during city driving conditions.

Highway MPG (Miles Per Gallon):The estimated fuel efficiency in miles per gallon during highway driving conditions.

Price:The cost of the vehicle.

Fuel Type (Gasoline): Indicates whether the vehicle is powered by gasoline.

Feature Selection – Techniques

- **Filter methods**
- **Wrapper methods**
- **Embedded methods**

1:Filtered Method:

Filter methods are typically employed as an initial preprocessing step in feature selection. Unlike wrapper methods, which integrate feature selection within machine learning algorithms, filter methods operate independently of any specific algorithm. Instead, features are chosen based on their scores derived from various statistical tests that assess their correlation with the target variable.The primary reliance of filter methods is on the inherent characteristics of the data, focusing on feature properties rather than involving machine learning algorithms directly. This characteristic makes them less computationally expensive compared to wrapper methods. However, it's noteworthy that their predictive performance tends to be lower than wrapper methods.

Filter methods are particularly well-suited for a rapid screening and removal of irrelevant features. Their efficiency lies in quickly identifying features that exhibit strong statistical relationships with the outcome variable. While they might not yield the highest prediction accuracy, their expeditious nature makes them valuable in scenarios where computational resources are limited or when a swift initial assessment of feature relevance is essential.

Basic Method :-

In our initial filtering step, we employed a fundamental technique to eliminate constant and quasi-constant features from the dataset. This process involved the application of a variance threshold, specifically set at 0.1. The variance threshold mechanism systematically identifies and removes features whose variance fails to meet the predefined threshold. In this context, features exhibiting minimal variation, whether constant or nearly constant, were excluded from the dataset. This approach serves as a foundational step in feature selection, paving the way for

subsequent analyses by focusing on variables that contribute meaningful variability to the dataset.

2: Wrapper Method:

In wrapper methods, the approach involves iteratively selecting subsets of features and training a model with each subset. The decision to include or exclude features from the subset is based on insights gained from the performance of the previous model. Essentially, this process transforms into a search problem, where the goal is to find the most informative combination of features for optimal model performance. However, it's important to note that wrapper methods tend to be computationally intensive due to the exhaustive search over feature subsets, making them resource-demanding.

Various wrapper methods are :

- Forward selection
- Backward elimination
- Exhaustive feature selection
- Recursive feature elimination
- Recursive feature elimination with cross-validation

In, our feature selection we have done the Forward selection method.

Forward Stepwise selection:

It is a method for variable selection in statistical modeling. The process initiates with a Null Model. Subsequently, the method systematically incorporates the most statistically significant variables into the model, one at a time. This iterative addition of variables continues until a predetermined stopping rule is met, or until all the variables under consideration have been included in the model. The primary objective is to refine the model by iteratively introducing the most relevant variables based on statistical significance, ultimately enhancing its predictive capabilities.

we determined the threshold value as 0.01. The most significant variable was choosen on the criteria that it has the smallest p-value compared to the given set threshold

value.Therefore, the features we got based on this forward selection process are 'engine_size', 'width', 'horse_power', 'stroke', 'fuel_type_gas', and 'peak_rpm'.

3: Embedded Methods :

Embedded methods follow an iterative approach during the model training process, systematically extracting features that contribute significantly to each iteration. Among the prevalent embedded methods, regularization stands out as a widely adopted technique. This method imposes penalties on features based on predefined coefficient thresholds.

Regularization, often referred to as a penalization method, introduces supplementary constraints into the optimization process of predictive algorithms, particularly in regression. These constraints are designed to guide the model toward lower complexity, favoring models with fewer coefficients. Notable instances of regularization methods include LASSO, RIDGE regression, and Elastic Net. These techniques incorporate built-in penalization functions aimed at mitigating overfitting, ultimately enhancing the model's generalizability and performance.

In our variable extracting we have done LASSO Regression.

LASSO Regression:

It stands for **L**east Absolute **S**hrinkage and **S**election **O**perator, a variant of linear regression, incorporates a technique known as shrinkage, wherein data values are pulled towards a central point, often the mean. This method is specifically designed to promote the development of uncomplicated, sparse models—models with fewer parameters. Lasso regression is particularly advantageous in scenarios characterized by high levels of multicollinearity or when there's a need to automate aspects of model selection, such as variable selection or parameter elimination.

The key mechanism in lasso regression involves L1 regularization, which imposes a penalty equivalent to the absolute value of the coefficients' magnitudes. Regularization, in general, entails introducing a penalty to the various parameters of a machine learning model to constrain its flexibility and, consequently, mitigate the risk of overfitting. In the context of linear models, this penalty is applied to the coefficients that scale each predictor.

Lasso's distinctive attribute within the spectrum of regularization techniques is its ability to shrink certain coefficients all the way to zero. Consequently, features associated with these zeroed-out coefficients can be effectively removed from the model. This property makes lasso

regression a powerful tool for feature selection, providing a means to streamline models by automatically identifying and excluding less influential predictors.

The features that were selected by the LASSO regression model based on their non-zero coefficients are : 'curb_weight', 'engine_size', 'compression', 'horse_power', and 'peak_rpm'.

Comparing the above Three features selection and the Model1 from HW2:

Interpreting the results of feature selection and comparing the selected features:

Common Features Selected by All Methods:

'curb_weight': The weight of the car when it's ready to drive.

'peak_rpm': The maximum revolutions per minute of the engine.

'comprassion': Compression ratio of the engine.

'horse_power': The horsepower of the engine.

'engine_size': The size of the car's engine.

These features are consistently identified as important across different feature selection methods, suggesting they have a significant impact on predicting the dependent variable 'price.'

Features with Non-Zero Coefficients in Model1:

'wheel_base': The distance between the centers of the front and rear wheels.

'length': The length of the car.

'width': The width of the car.

'height': The height of the car.

'curb_weight': The weight of the car when it's ready to drive.

'engine_size': The size of the car's engine.

'bore': The diameter of the engine cylinders.

'stroke': The length of the engine's pistons moving up and down.

'compression': Compression ratio of the engine.

'horse_power': The horsepower of the engine.

'peak_rpm': The maximum revolutions per minute of the engine.

'city_mpg': Miles per gallon in the city.

'highway_mpg': Miles per gallon on the highway.

'fuel_type_gas': Binary indicator for gas fuel type.

These features with non-zero coefficients in the full linear regression model (model1) are considered significant in predicting 'price' based on their contribution to the linear regression equation.

The common features selected by all methods and the features with non-zero coefficients in model1 represent aspects of the car that strongly influence its price.

'engine_size', 'curb_weight', 'horse_power', and 'peak_rpm' appear to be consistently important across all methods, indicating the significant role of the car's engine characteristics.

Other features like 'wheel_base', 'length', 'width', 'height', 'bore', 'stroke', 'compression', 'city_mpg', 'highway_mpg', and 'fuel_type_gas' also contribute to predicting the 'price' in the linear regression model.

Principle Components Analysis:

Principal Component Analysis (PCA) is a dimensionality reduction technique widely used in statistics, machine learning, and data analysis. The primary objective of PCA is to transform a high-dimensional dataset into a lower-dimensional one, capturing the essential information while minimizing information loss.

Key points about PCA:

Dimensionality Reduction: PCA is employed to reduce the number of features (or dimensions) in a dataset while retaining as much of the original variability as possible. This is especially valuable in datasets with a large number of correlated variables.

Orthogonal Transformation: PCA performs an orthogonal linear transformation to convert the original variables into a new set of uncorrelated variables called principal components. These components are ordered by the amount of variance they explain in the data.

Variance Maximization: The first principal component explains the maximum variance in the data, and each subsequent component captures as much of the remaining variance as possible. This allows for a concise representation of the dataset with fewer components.

Eigenvalues and Eigenvectors: PCA involves the computation of eigenvalues and eigenvectors of the covariance matrix of the original data. The eigenvectors represent the directions of maximum variance, and the corresponding eigenvalues indicate the magnitude of variance along these directions.

In our Dataframe with reduced features is obtained through Principal Component Analysis (PCA). Each row represents an observation from our original dataset. The columns 'PC1' through 'PC5' are the principal components resulting from the PCA transformation, and the 'price' is the dependent variable.

Each principal component (PC) is a linear combination of the original features.

These principal components are orthogonal to each other, and they capture the maximum variance in the data.

The values in each row of the 'PC1' to 'PC5' columns represent the coordinates of the data points in the reduced feature space.

PC1, PC2, and PC3 have values that vary across different observations, indicating variations in the dataset along these directions.

PC4 and PC5, being orthogonal, capture additional variations in the data not covered by the first three components.

Price Column:

The 'price' column represents the dependent variable.

Each row corresponds to the price of the item associated with the feature values represented by the principal components.

Overall Interpretation:

The reduced feature space represented by 'PC1' to 'PC5' condenses the information from the original features while retaining the most significant variations.

The 'price' column allows you to associate the reduced feature values with the original target variable, facilitating analysis or modeling with a lower-dimensional representation.

Regression:

After doing the PCA we performed a regression analysis on the new dataframe some of our findings were :

In Ordinary Least Squares (OLS) regression results, several key findings are :

The dependent variable in the regression analysis is 'price.'

The overall fit of the model is described by the R-squared value, which is 0.236. This implies that approximately 23.6% of the variability in the 'price' variable is explained by the independent variables (PC1, PC2, PC3, PC4, PC5).

The adjusted R-squared, accounting for the number of predictors, is 0.215.

Variable Coefficients:

The intercept (const) is estimated at 1.304e+04, and its t-statistic is highly significant ($t = 25.061$, $p < 0.0001$).

Among the principal components (PC1 to PC5), only PC1 has a statistically significant coefficient with a positive impact on 'price' (coefficient = 1445.4747, $t = 7.344$, $p < 0.0001$).

Statistical Significance:

The F-statistic, assessing the overall significance of the regression model, is 11.32, and the associated p-value is very low (Prob (F-statistic) = 1.60e-09), indicating that the model is statistically significant.

The p-values for individual coefficients provide insights into their significance. For instance, PC2, PC3, PC4, and PC5 are not statistically significant at the 0.05 level.

Model Diagnostics:

The Omnibus test and Jarque-Bera test suggest potential issues with the normality of residuals.

The Durbin-Watson statistic is 0.471, indicating the presence of positive autocorrelation.

The Cond. No. of 3.30 suggests potential multicollinearity.

Recommendations:

Considering the statistical significance and impact on 'price,' further investigation into the relationship with PC1 is warranted.

Attention should be given to addressing potential issues with normality, autocorrelation, and multicollinearity to enhance the robustness of the model.

In summary, the model demonstrates overall significance and provides insights into the impact of certain principal components on 'price'.

Conclusion

In conclusion, our feature selection process employed three distinct techniques: Filter methods, Wrapper methods, and Embedded methods, each revealing valuable insights into the most influential features for predicting the 'price' of automobiles.

Filter methods, such as the variance threshold applied in our initial preprocessing step, efficiently screened and removed irrelevant features based on their statistical characteristics. Although less computationally expensive, filter methods may sacrifice some predictive performance compared to more intricate wrapper methods.

Wrapper methods, exemplified by Forward Stepwise selection, engaged in an iterative approach, selecting subsets of features based on their impact on model performance. Despite their computational intensity, wrapper methods offer a more exhaustive exploration of feature subsets, potentially leading to higher predictive accuracy.

Embedded methods, specifically LASSO Regression, showcased the power of regularization in extracting significant features. By imposing penalties on coefficients, LASSO efficiently performed feature selection, identifying 'curb_weight,' 'engine_size,' 'compression,' 'horse_power,' and 'peak_rpm' as crucial contributors to predicting 'price.'

Principal Component Analysis (PCA) further reduced dimensionality, condensing the dataset into five principal components. The subsequent regression analysis revealed that PC1 significantly impacted 'price,' emphasizing the importance of these orthogonal components in capturing variance within the data.

Overall, our findings underscore the significance of features such as 'engine_size,' 'curb_weight,' 'horse_power,' and 'peak_rpm' in determining automobile prices. The diverse approaches of filter, wrapper, and embedded methods, along with PCA, provided a comprehensive understanding of feature relevance. As we move forward, addressing issues related to normality, autocorrelation, and multicollinearity identified in the regression analysis will be crucial for refining the model's robustness and enhancing its predictive capabilities.