DATA SCIENCE ASSIGNMENT-03

NUTHAN REDDY VADDI REDDY

UTA ID: 1002175133

---

1. Data from our lives:

Consider the Employee's salary dataset of having multiple features.

Feature 1: Years of experience.

Feature 2: level of Education.

Feature 3: Location.

Feature 4: Job role or position.

Feature 5: Industry or sector of employment.

Feature 6: Age

Feature 7: gender

Feature 8: Employee id

Feature 1: Years of experience:

It is one of the important factors for predicting the employee salary. How many years of real-time experience that the person has will lead to a high salary, as an experienced person will have more real-time experience and skill on the work.

Feature 2: level of Education:

High level of education degrees may lead to higher salaries like the ph.d student or can get higher salary than an undergraduate student, By considering both the level of education and the field of study.

Feature 3: Location:

The cost of living and taxes varies by different locations. Salaries in areas like California will tend to increase higher for increased living expenses(cost of living) and high taxes. so, a person in California will get a higher salary than a person in Texas, so the location is an important predictor in the employee salary.

Feature 4: Job role or position:

Different roles will have different salaries within the same organization.For example, a Data scientist will earn more than a data analyst.

Feature 5: Industry or sector of employment:

The industry or sector in which the employee works can have an impact on salary. For example, salaries in the technology sector might be higher than those in finance.

Feature 6: Age:

The age of the person will not impact the person's salary. so,can remove the feature age to train the data to avoid the overfitting and Improve accuracy.

Feature 7: Gender:

The Gender of the person will not impact one person's salary. so,can remove the feature Gender to train the data to avoid the overfitting and Improve accuracy.

Feature 8: Employee id:

The Employee id of the person will not impact on the person's salary. so,can remove the feature Employee id to train the data to avoid the overfitting and Improve accuracy.

So, The Employee id,gender and age will not be an important feature to classify the Employee salary. So, We can remove or drop these features. Years of experience,level of Education,Location,Job role or position and Industry or sector of employment are important features to calculate Employee salary.

Consider the student's dataset of a school having multiple features.

Feature 1: Student's Roll number

Feature 2: Weight of the student

Feature 3: Height of the student

Feature 4: gender of the student

Student's roll number and gender is not an important feature to classify whether a person is obese or not. It has nothing to do with the physical state of a person. So, We can remove or drop these features. Height and weight are important features to calculate BMI and classify students. So, we select Height and Weight as our features.

2. Variable selection:

2.1. Filtered methods:

- I am using the correlation heatmap method in filtered methods for feature selection:
- The features are selected on the basis of their scores in various statistical tests for their correlation with the outcome variable and removing the features which are highly correlated with each other and taking only one feature out of highly correlated features for training.
- From the total features the wheel_base,heights,bore,stroke,comprassion,horse_power,peak_rpm and city_mpg features are selected.
- The length is highly correlated(above threshold) with wheel_base and width, so we are taking wheel_base for training and removing length and width from the dataset.
- comprassion is highly correlated(above threshold) with engine_size and fuel_type_gas, so we taking comprassion for training and removing engine_size and fuel_type_gas from the dataset.

- city_mpg is highly correlated(above threshold) with highway_mpg, so we are taking city_mpg for training and removing engine_size from the dataset.
- The heights,bore and stroke are not correlated with any of the features above threshold so should keep the feature for training dataset.

2.2 Wrapper methods:

Recursive Feature Elimination:

- Recursive Feature Elimination (RFE) is a feature selection to eliminate the least important features until the desired number of features is reached and Build a model using the remaining features.
- Initializes the Recursive Feature Elimination (RFE) object with the SVM classifier as the estimator. we specify that it should select 7 features (n_features_to_select=7) in each iteration with a step size of 1. The RFE is then fit to the training data.
- From Recursive Feature Elimination (RFE) Selected Features are length,width,peak_rpm,heights, curb_weight,engine_size and horse_power.
- The least important features are wheel_base,bore,stroke,comprassion,city_mpg,highway_mpg and fuel_type_gas.

Forward selection

- In feature selection first we are taking the list in the starting

- then starts adding the most significant variables one after the other, Until a pre-specified stopping rule is reached or until all the variables under consideration are included in the model.
- after adding first variable the Mean R-squared: 0.7604677398744579
- after adding all the variables Mean R-squared: 0.7604677398744579
- Based on the analysis and visualizations conducted, it is evident that opting for 8 features results in the highest mean R-squared value. This finding suggests that selecting 8 features strikes an optimal balance between model complexity and predictive performance. Consequently, we have determined the number of features to be selected as 8 for our model.
- The features chosen through the Wrapper Method (Forward Selection) comprise 'length', 'width', 'heights', 'curb_weight', 'engine_size', 'stroke', 'peak_rpm', and 'fuel_type_gas'.

2.3. Embedded methods:

- In the lasso regression with the GridSearchCV and KFold.
- In Lasso regression performs feature selection by setting some coefficients to exactly zero. The specific features selected depend on the value of the regularization parameter alpha.
- The GridSearchCV is used to perform a cross-validated grid search over a specified range of alpha values (from 0.00001 to 10 with a step of 500). It uses 5-fold cross-validation (KFold) for evaluating each combination of hyperparameters.

- The best hyperparameter (alpha) is selected based on the mean cross-validated score.
- After obtaining the best alpha from the grid search, a new Lasso model (lasso1) is created with that best alpha.
- This model is then fitted to the training data (X_train and y_train).
- The absolute values of the coefficients obtained from the fitted Lasso model (lasso1) are calculated and Features with absolute coefficients greater than 60 are selected as the chosen subset of features.
- The features are selected :
- length,width,heights,engine_size,bore,stroke,city_mpg,highway_mpg,fuel_type_gas by selecting the absolute coefficients greater than 60**,** we can change the threshold based on your requirement.

2.4. Compare your results:

Model 1 is constructed using all 14 features. The R-squared value, which is 0.860, indicates that this model accounts for 86% of the variability in the dependent variable 'price'.

coefficients:

const 1.325e+04

wheel_base 242.4026

length -756.4877

width 1287.2555

heights 789.8987

curb_weight 619.0835

engine_size 5736.6775

bore -328.5262

stroke -1164.1252

comprassion -2506.9293

horse_power 1311.5357

peak_rpm 1194.8754

city_mpg -1845.4337

highway_mpg 2162.3895

fuel_type_gas -3568.8242

Model 2 is built using the features selected using the Filtered Method Correlation. The R-squared value is 0.778 indicates that the model can explain about 84.9 % of the variation in the dependent variable.

Wheel_base coefficients is increased in model 2 after feature selection

Heights coefficients is decreased in model 2 after feature selection

Engine size coefficients is increased in model 2 after feature selection

Bore coefficients is increased in model 2 after feature selection

Stroke coefficients is increased in model 2 after feature selection

Compression coefficients is decreased in mode 2 after feature selection

Horse_power coefficients is increased in model 2 after feature selection

Peak_rpm coefficients is decreased in model 2 after feature selection

City_mpg coefficients is increased in model 2 after feature selection

coefficients:

const 1.325e+04

wheel_base 728.3277

heights 559.3257

engine_size 5528.4133

bore -313.6890

stroke -892.6093

comprassion 1207.6088

horse_power 1730.9722

peak_rpm 951.8232

city_mpg -412.2518

Model 3 is constructed with features selected through the wrapper method, specifically Forward Selection. The R-squared value, standing at 0.85, signifies that this model has the capability to elucidate approximately 85% of the variability in the dependent variable.

Length coefficients is increased in model 3 after feature selection

Width coefficients is decreased in model 3 after feature selection

Heights coefficient is decreased in model 3 after feature selection.

Curb_weight coefficients is increased in model 3 after feature selection

Engine_size coefficients is increased in model 3 after feature selection

Stroke coefficients is increased in model 3 after feature selection

Peak_rpm is decreased in model 3 after feature selection

coefficients:

const 1.325e+04

length -706.7578

width 1205.1433

heights 597.5580

curb_weight 1260.5564

engine_size 6129.7443

stroke -985.6500

peak_rpm 1457.1932

fuel_type_gas -849.3803

Model 4 is created utilizing features selected through the embedded method, specifically Random Forest. The R-squared value, recorded at 0.715, suggests that this model is proficient in clarifying around 71.5% of the variability in the dependent variable.

Length coefficients is increased in model 4 after feature selection

Heights coefficients is decreased in model 4 after feature selection

Curb_weight coefficients is increased in model 4 after feature selection

coefficients:

const 1.325e+04

length -578.0126

heights -914.4132

curb_weight 7523.9189

Comparison:

The coefficients from the full linear regression model (model1) are the original coefficients based on all features.
correlation,RFE and Lasso selected a subset of features, and their coefficients are specific to the selected features.
While some coefficients are similar across the methods, differences exist due to the different strategies each method employs for feature selection.
Feature selection aims to identify the most relevant features, resulting in adjusted coefficients for the selected features in each method.

Model 1 boasts the highest R-squared value at 0.860, signifying its capacity to account for the most variation in the dependent variable (price). Model 2 follows closely with a slightly lower R-squared value of 0.849, while Model 3 demonstrates an R-squared value of 0.850 and Model 4, however, exhibits the lowest R-squared value at 0.715.

Given these findings, Model 1, which employs all features, is discarded due to its utilization of the maximum number of features. Among the remaining three models, Model 3 emerges as a potentially more reliable choice, showcasing a commendable R-squared value with a relatively smaller number of features.

The coefficients across the four models exhibit similarities in both direction and magnitude. Nevertheless, notable distinctions are observed. For example, the coefficient corresponding to the 'wheel_base' variable is consistently positive and statistically significant in all models. Notably, in Model 1, it assumes a larger value of 39.53 in the other models. This discrepancy implies that in Model 1, 'wheel_base' holds a more pronounced positive correlation with 'price' compared to the other models.

Another noteworthy distinction emerges in the coefficient associated with the 'fuel_type_gas' variable. While this variable is consistently negative and statistically significant across all four models, its magnitude stands out in Model 3, registering at -849.3803. In contrast, the other models exhibit comparatively smaller magnitudes. This suggests that in Model 3, 'fuel_type_gas' bears a more pronounced negative relationship with 'price' compared to the other models.

In general, the coefficients across the four models demonstrate a consistent pattern. Nevertheless, discernible differences between the models highlight that the association between 'price' and the independent variables is not uniformly linear. These variations imply that the nature of this relationship might be influenced by the specific values assumed by other independent variables in the models.

PCA:

Based on the visual analysis of the explained variance ratio graph, we can deduce that utilizing 8 principal components is sufficient to capture over 97.4% of the variance in the data. Consequently, we can confidently choose 8 components for our regression analysis

The PCA model successfully captures 84.6% of the variance, a performance comparable to the models constructed using all features and Forward Selection. This highlights the efficacy of leveraging principal components, achieving a balance between model simplicity and performance. By effectively capturing the variability in car prices, the PCA-based approach demonstrates its ability to represent the essential information within the dataset while reducing dimensionality.